NCEE 2012-4017

U. S. DEPARTMENT OF EDUCATION

Effects of the Connected Mathematics Project 2 (CMP2) on the Mathematics Achievement of Grade 6 Students in the Mid-Atlantic Region

NATIONAL CENTER FOR EDUCATION EVALUATION AND REGIONAL ASSISTANCE
Institute of Education Sciences

Effects of the Connected Mathematics Project 2 (CMP2) on the Mathematics Achievement of Grade 6 Students in the Mid-Atlantic Region

Final Report

March 2012

Authors:

Taylor Martin, The University of Texas at Austin

Sarah J. Brasiel, Edvance Research, Inc.

Herb Turner, ANALYTICA, Inc.

John C. Wise, The Pennsylvania State University

Project Officer:

OK-Choon Park
Institute of Education Sciences

NCEE 2012-4017

U.S. Department of Education

U.S. Department of Education

Arne Duncan

Secretary

Institute of Education Sciences

John Q. Easton

Director

National Center for Education Evaluation and Regional Assistance

Rebecca A. Maynard

Commissioner

March 2012

This report was prepared for the National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, under contract ED-06CO-0029 with Regional Educational Laboratory Mid-Atlantic administered by the Pennsylvania State University.

IES evaluation reports present objective information on the conditions of implementation and impacts of the programs being evaluated. IES evaluation reports do not include conclusions or recommendations or views with regard to actions policymakers or practitioners should take in light of the findings in the report.

This report is available on the Institute of Education Sciences website at http://ncee.ed.gov and the Regional Educational Laboratory Program website at http://edlabs.ed.gov.

Alternate Formats Upon request, this report is available in alternate formats, such as Braille, large print, audiotape, or computer diskette. For more information, please contact the Department’s Alternate Format Center at 202-260-9895 or 202-205-8113.

Disclosure of potential conflict of interest

None of the authors or other staff involved in the study from the Regional Educational Laboratory Mid-Atlantic, The Pennsylvania State University, ANALYTICA, Inc., Rutgers Center for Effective Practices, ICF International, Metiri Group, The University of Texas at Austin, and Edvance Research, has financial interests that could be affected by the content of this report. No one on the five-member Technical Working Group, convened twice annually by the research team to provide advice and guidance, has financial interests that could be affected by the study findings.[1]

[1] Contractors carrying out research and evaluation projects for IES frequently need to obtain expert advice and technical assistance from individuals and entities whose other professional work may not be entirely independent of or separable from the tasks they are carrying out for the IES contractor. Contractors endeavor not to put such individuals or entities in positions in which they could bias the analysis and reporting of results, and their potential conflicts of interest are disclosed.

Contents

Tables

Figures

Executive summary

Mathematics proficiency is critical for student success (Bailey 2009; Grubb and Cox 2005), access to postsecondary education (Greene and Forster 2003), and preparation for future employment (Bishop 1988). In the United States, middle school mathematics is a gateway to mathematics in high school (Useem 1992), required for acceptance into many four-year colleges (Adelman 2006). Students' success in mathematics in high school and college is associated with better employment prospects (Steen 2007) and more money earned over their lifetimes (Betts and Rose 2001).

Mathematical literacy is a growing need in our increasingly technological society (Meaney 2007). The National Research Council and the American Association for the Advancement of Science stress the importance of improving mathematics instruction and achievement to improve students' abilities—not just in mathematics but in science, technology, and engineering as well (Augustine, Vagelos, and Wulf 2005).

This study examines the effects of Connected Mathematics Project 2 (CMP2) on grade 6 student mathematics achievement and engagement using a cluster randomized controlled trial (RCT) design. It responds to a need to improve mathematics learning in the Mid-Atlantic Region (Delaware, Maryland, New Jersey, Pennsylvania, and Washington, DC).

Connected Mathematics Project 2

At the time of this study, CMP2 was the latest version of the Connected Mathematics Project (CMP). Designed for use in grades 6–8, CMP2 allows students to be responsible for their learning by exploring different solution pathways, sharing their ideas with other students, listening to the ideas of others, and questioning each other. Teachers ensure that the mathematics goals of the lesson are addressed and that students develop conceptual understanding and procedural skills, by asking them questions and encouraging them to share their thinking, compare their thinking with others, and make connections between representations of problems and solutions.

CMP2 incorporates elements identified in policy reports and research articles as key components of effective mathematics instruction:

- Active student participation in building knowledge through making conjectures, justifying solutions, and clarifying ideas orally and in writing (National Council of Teachers of Mathematics 1991; Elbers 2003).
- Motivation and engagement of students through the use of contexts connected to their experiences and background (Chung 2004; National Council of Teachers of Mathematics 1989, 1991, 2000; Powell 2006; Ridlon 2004; Tsao 2004).
- Collaborative learning through group problem solving activities (Cobb 1989; Erlwanger 1973; National Council of Teachers of Mathematics 1991; Schwab 1975).
- The development of conceptual understanding and procedural knowledge (Battista 2001; Cobb, Yackel, and Wood 1992; Lucas 2006; Popkewitz 2004).

The teacher professional development (PD) for CMP2 is intended to deepen teacher understanding of mathematics, strengthen their pedagogical knowledge, and develop their ability to use inquiry-based instructional strategies. It also exposes them to a variety of methods of assessing student understanding and progress and using assessment results to inform instructional decisions.

When using CMP2, students progress through a series of units made up of problem-centered investigations that address core mathematics topics. In each investigation, students are introduced to a problem, work in groups to share ideas about the problem and to develop solution strategies, and justify their thinking by presenting their solution strategies to the class. Students can practice their understanding and skills in additional problem sets included with each investigation.

The teacher facilitates the lesson by introducing the problem and connecting new content to prior understanding and investigations. He or she circulates around the class—asking questions and helping students when needed during the open work period—and helps connect presentations to prior learning and core mathematical concepts during the final presentation phase. This is in contrast to the finding by Stigler et al. (1999) that 90 percent of the class time in observed United States mathematics classes is spent on practicing routine procedures.

Research base for CMP/CMP2

In 2005, the What Works Clearinghouse (WWC)[2] reviewed 22 studies that examined the effects of CMP[3] on middle school student mathematics outcomes. None of the studies met WWC standards[4] for experimental or quasi-experimental studies. Three quasi-experimental studies (Ridgway et al. 2003; Riordan and Noyce 2001; Schneider 2000) qualified as meeting standards “with reservations” (What Works Clearinghouse 2005). Results from these studies are mixed. Schneider (2000) found that CMP students scored lower than the comparison group on the outcome measure; Riordan and Noyce (2001) found that CMP students scored higher than a comparison group; and Ridgway et al. (2003) found results that varied across outcome measures and grade levels.[5]

In 2010, the WWC updated its review of research on CMP[6] (What Works Clearinghouse 2010). An additional 57 studies were identified, bringing the total to 79. Using more stringent criteria than the previous review, however, only Schneider (2000) still met WWC evidence standards “with reservations” (What Works Clearinghouse 2010). Ridgway et al. (2003) and Riordan and Noyce (2001) failed to establish baseline equivalence between their groups.

[2] The What Works Clearinghouse was established in 2002 by the Institute of Education Sciences to provide educators, policymakers, researchers and the public with a central and trusted source of scientific evidence of what works in education. http://ies.ed.gov/ncee/wwc/
[3] CMP2 is the revised version of the Connected Math Program (CMP) and was published in 2006 (Lappan et al. 2006)
[4] WWC categorizes studies as Meets Evidence Standards, Meets Evidence Standards with Reservations (highest possible for quasi-experimental studies), and Does Not Meet Evidence Screens. A complete description of these categories is available at http://ies.ed.gov/ncee/wwc/help/idocviewer/doc.aspx?docid=20&tocid=1
[5] Grade 6 and 7 CMP students scored lower than a comparison group on the Iowa Test of Basic Skills (ITBS), while grade 8 students scored higher. On the Balanced Assessment of Mathematics (BAM), grade 6 through 8 CMP students scored higher than a comparison group.
[6] The 2010 WWC review does not differentiate between CMP and CMP2.

While the current study was in progress, an RCT examining CMP2 was published (Eddy et al. 2008). It was not included in the 2010 WWC review and reported no statistically significant effects.

To summarize, one quasi-experimental (Schneider 2000) and one experimental (Eddy et al. 2008) study failed to identify a statistically significant impact on mathematics achievement for either CMP or CMP2. Other quasi-experimental studies originally reported statistically significant effects but were later found to not meet WWC evidence standards (Riordan and Noyce 2001; Ridgway et al. 2000). To investigate the effectiveness of CMP2, a well-designed experimental study is needed.

Current study

The current study is a cluster RCT designed to evaluate the effect of CMP2 on the mathematics achievement of grade 6 students. Schools were the unit of random assignment in this study, and CMP2 was implemented at the school level. The study spanned two years:

- An implementation year, in which teachers were trained in CMP2 and began using it in the classroom (2008/09).
- The focal year for the impact evaluation, which we refer to throughout as the "impact year," in which teachers implemented CMP2 and student-level achievement data were collected (2009/10).

Research questions

This study addresses the following primary research question:

What is the impact of being in a school randomly assigned to adopt CMP2 on grade 6 student mathematics achievement?

A statistically significant difference in outcomes in favor of CMP2 would support the use of CMP2.

Research over the last two decades has shown a relationship between achievement and school-related variables such as academic engagement, perceptions and attitudes, self-confidence in learning mathematics and science, interest and motivation to learn mathematics and science, and self-efficacy (Eccles and Jacobs 1986; Helmke 1989; Reynolds and Walberg 1991, 1992). Therefore, the study included a secondary research question:

- What is the impact of being in a school randomly assigned to adopt CMP2 on the value that grade 6 students place on mathematics?

The answer to this secondary question was not intended to determine the effect of CMP2 on mathematics achievement but to supplement the answer to the primary question—by examining whether CMP2 had an effect on the psychosocial outcome of perceived task value (PTV). The answer to the secondary research question alone would be insufficient for assessing the impact of CMP2.

Measures

The TerraNova CAT™2 Basic Multiple Assessments Form (CTB/McGraw-Hill 2003) was used as both the baseline and outcome measure of mathematics achievement in the study. Students were given the standard 90 minutes to complete the test. The TerraNova covers the following mathematics concepts: number and number relations; computation and numerical estimation; operation concepts; measurement; geometry, and spatial sense; data analysis, statistics, and probability; patterns, functions, and algebra; problem solving and reasoning; and communication. The TerraNova has a reliability of 0.91 based on a nationally representative sample (McGraw-Hill 2002).

The PTV, a subscale of a validated mathematics engagement survey (Eccles and Wigfield 1995), was used as both the baseline and outcome measure of the value that students place on mathematics. On this measure, students describe their perceptions of the attractiveness, importance, and utility of a particular task or content.

The study team also collected school, teacher, and student data as covariates for the analyses and used data from classroom observations to help understand program implementation. School data were drawn from the Common Core of Data (CCD; NCES n.d.b) and included school-level demographics such as locale, student population by subgroup, and average percentage eligible for free or reduced-price lunch. Teacher data included a background survey, monthly online surveys reporting on progress using CMP2, and an end-of-year survey summarizing time spent weekly on mathematics and the CMP2 units completed during the school year. Student data were pretest scores on the TerraNova and PTV measures. Classrooms were observed twice during the impact year to examine implementation. These data included the length of the class period, allocation of time to different activities (including those seen as more or less similar to activities promoted by CMP2), and observed teacher and student practices.

Study sample and schedule

Prior to recruitment (2008), the CCD (NCES n.d.a) was used to identify all public and charter schools enrolling grade 6 students in the Mid-Atlantic region. Beginning January 2008, invitations were sent to 989 districts comprising 2,597 schools. The incentive for schools randomly assigned to CMP2 included free curriculum materials for teachers and students and free PD (including trainer fees, teacher stipends, substitute costs, and transportation costs) for both study years. Control schools received $1,000 for participating.

A total of 105 schools across 80 districts submitted letters of interest. Of these, 73 schools had not previously used CMP or CMP2 and expressed a willingness to abide by the guidelines for participating in the study. These schools were invited to sign a memorandum of understanding in March 2008. The number fell to 70 before randomization due to schools merging or withdrawing.[7]

In May 2008, the 70 schools were randomly assigned within jurisdiction[8] to study conditions, 36 to the intervention group and 34 to the control group. The imbalance in group size was due to

[7] Greater detail cannot be provided in order to prevent disclosure of the schools involved.
[8] Jurisdictions in the Mid-Atlantic region include Delaware, Maryland, New Jersey, Pennsylvania, and Washington, DC. Four of these jurisdictions had schools that participated in this study.

chance. Between the point of random assignment and the start of the impact year (2009/10), five schools were lost due to school-level administrator decisions to withdraw schools from the study and district-level decisions to close and merge campuses with low enrollment. The schools that dropped out or merged were not statistically significantly different from the remaining sample of 65 schools on any of the measured baseline school characteristics.

During the implementation year (2008/09), intervention teachers were offered the “typical” CMP2 PD administered by the publisher (two days before the school year and three days during; M. Baughman, personal communication, January 2007) and received all standard curriculum materials. Teachers in the control schools continued to use their schools’ mathematics curricula. The purpose of the implementation year was to give intervention teachers time to become accustomed to delivering a new curriculum; no student performance data were collected during the implementation year.

During the impact year (2009/10), intervention schools continued using the standard CMP2 curriculum materials they had received during the implementation year. New teachers in intervention schools were again offered the standard PD administered by the publisher. Control teachers continued using their respective school’s regular curriculum (business as usual). TerraNova and PTV data were collected at the beginning (pretest) and end (posttest) of the school year.

Analysis and results

The final analysis included 65 schools, including 5,677 students for the TerraNova and 5,584 for the PTV. This was 82 percent of the eligible students (students enrolled in a regular grade 6 mathematics class in a study school at the time of pretest) for the TerraNova at posttest and 80 percent of the eligible students for the PTV at pretest.[9] A two-level hierarchical linear model of students nested within schools was used to estimate the impact of CMP2 on the primary and secondary student outcomes. To improve the estimates, covariates such as school locale that were identified as statistically significantly different at baseline were included in the model at level 2, except for student pretest scores, which were included at both levels. Sensitivity analyses were conducted to test the robustness of the results under alternative model specifications.

The impact of CMP2 on student TerraNova posttest scores was less than one point (0.60), and was not statistically significant (effect size = 0.02; $p = .777$). Results for the secondary research question indicate that CMP2 was also no more effective than business as usual in improving students’ mathematics PTV (effect size = 0.09; $p= .109$). Sensitivity analyses found no changes in the direction or magnitude of the intervention effects.[10] A lack of correlation between pretest and posttest scores on PTV could indicate problems with either the measure or the underlying construct itself. The secondary research question findings should thus be interpreted with caution.

[9] Eligible students include 761 students no longer in the study due to the merger or withdrawal of their schools prior to pretesting. These students represent 11 percent of the eligible students and were counted as attrition.
[10] A sensitivity analysis conducted without covariates showed a statistically significant difference in TerraNova outcomes in favor of the control group, but this finding can be explained by differences in pretest scores measured at baseline.

Data from the spring 2010 (year 2) classroom observations showed statistically significant differences between the intervention and control schools on all six measured variables,[11] indicating a contrast in instruction between the two groups. In particular, the percentage of class time dedicated to activities seen as more like those promoted by CMP2 was statistically significantly higher in intervention schools (34 percent; $p = .000$). More intervention teachers were observed engaging in behaviors intended to foster student responsibility for learning and complex thinking (difference of 3.24 of 11 points, $p = .000$), and more intervention students were observed demonstrating responsibility for learning and complex thinking in class discussion (a difference of 1.00 of 5 points, $p = .004$) and in groups or pairs (a difference of 2.80 of 5 points, $p = .000$).

Data from teacher self-reports showed that 68 percent of intervention teachers met the publishers' recommended 50 minutes per day and 64 percent completed the recommended six units per school year. There was also a statistically significant difference between intervention and control groups in the amount of time teachers reported spending on math, with intervention teachers spending an average of 1.18 more hours per week ($p = .002$).

Conclusions

The type of instructional activity taking place in intervention schools differed from that in control schools, and the activity observed in intervention schools was the type expected when implementing CMP2. Sixty-four percent of intervention teachers reported implementing the curriculum at a level consistent with the publishers' recommendations on the number of units completed per school year (six), and 68 percent of them reported implementing the curriculum consistent with the recommended amount of class time per week.

But CMP2 did not have a statistically significant effect on grade 6 mathematics achievement as measured by the TerraNova, which answered the primary research question.[12] Indeed, grade 6 mathematics students in schools using CMP2 performed no better or worse on a standardized mathematics test than did their peers in schools not using it. The results for the secondary research question were similar. There was no statistically significant difference between groups in PTV, and the small effect size is unlikely meaningful. These results were insensitive to alternative model specifications.

The lack of statistically significant effects is consistent with prior research on CMP2 rated in the 2010 WWC review as meeting standards "with reservations" (Schneider 2000) and the Eddy et al. (2008) RCT. The intent-to-treat analytical approach used in this study, which analyzes participants based on how they are randomly assigned, yielded unbiased estimates of program effectiveness as implemented.

To estimate the effect of CMP2 under typical conditions, teachers were provided all the typical materials and PD that a normal school adopting CMP2 would have. However, while CMP2 use

[11] These six variables include making connections, teacher factors related to student responsibility for learning and complex thinking, student evidence of responsibility for learning and complex thinking in class discussion, student evidence of responsibility for learning and complex thinking in groups/pairs, the percentage of class time spent on activities more like CMP2, and the percentage of class time spent on activities less like CMP2.
[12] The primary research question was designed to test the impact of CMP2 on mathematics achievement. The secondary research question was exploratory. Thus, no adjustment for multiple comparisons was performed.

was tracked, the study team did not ensure a particular amount or quality of CMP2 instruction. So, the curriculum impact reflects the effect of a school being assigned to use CMP2 or to continue use of their regular curriculum, not necessarily of actually using CMP2.

The results apply to the implementation of the CMP2 curriculum, after typical PD, in schools with grade 6 students. Use of a volunteer sample limits the findings to the schools, teachers, and students that participated in the study in the Mid-Atlantic region. The conclusions drawn in this study about the effects of CMP2 on student math achievement are limited to student math achievement as measured by the TerraNova, and do not generalize to any other standardized test.

1. Study Background

This chapter establishes the need for the current study and describes the intervention, including its development history, the reasons for its selection as the intervention for this study, and a review of previous studies of the intervention. The chapter concludes with a presentation of the theory of change and the counterfactual, a description of the study and research questions, and a summary of forthcoming chapters.

Need for the study

Importance of mathematics

Mathematics proficiency is critical for student success (Bailey 2009; Grubb and Cox 2005), access to postsecondary education (Greene and Forster 2003), and preparation for future employment (Bishop 1988). In the United States, middle school mathematics is a gateway to mathematics in high school (Useem 1992), required for entry into many four-year colleges (Adelman 2006). Students' success in mathematics in high school and college is associated with better employment prospects (Steen 2007) and more money earned over their lifetimes (Betts and Rose 2001).

Increased federal attention to mathematics

The National Commission on Excellence in Education's (NCEE 1983) report *A Nation at Risk* brought increased attention to improving mathematics education in the United States (Business-Higher Education Forum 2007). This attention led to defining standards for mathematics learning and legislating adequate progress toward attaining those standards. Standards for mathematics are defined for each state, and federal legislation—such as the Goals 2000: Educate America Act (P.L. 103-227) (1994) and the No Child Left Behind Act (P.L. 107-110) (2001)—have supported increased accountability for results. States are now required to plan how adequate yearly progress will be measured to determine the achievement of each school and school district and to hold them accountable for student performance, including achievement in mathematics (Elledge et al. 2009).

Math proficiency

Two large-scale assessments illustrate U.S. student performance in mathematics.[13] The first is the Trends in International Mathematics and Science Study (TIMSS). The TIMSS measures students' ability to solve mathematical and scientific problems through both multiple-choice and open-ended questions. Results from the TIMSS revealed that U.S. student mathematics performance in grades 4 and 8 has improved since the first TIMSS administration in 1995 but continues to lag behind other developed countries (Mullis, Martin, and Foy 2008).[14] The second is the National Assessment of Educational Progress (NAEP) mathematics assessment, designed to measure student understanding and proficiency in several core mathematics competencies: number properties and operations; measurement; geometry; data analysis, statistics, and probability; and algebra. Three types of questions are included in the NAEP mathematics

[13] The Program for International Student Assessment also provides data for international comparisons, but it only tests students older than 15 years and 3 months, so it is not included in this study of grade 6 students.
[14] The U.S. lags behind eight countries in Asia and Europe in grade 4 and behind five countries in Asia in grade 8.

assessment: multiple choice, short constructed response, and extended constructed response. The NAEP mathematics assessment is given in grades 4 and 8 to a nationally representative sample of students. In general, scores have been increasing since 1990. However, from 2007 to 2009, grade 4 student scores remained the same while grade 8 student scores continued to improve (Lee, Grigg, and Dion 2007; NCES 2009).

Although the TIMMS results showed improvement in mathematics for grades 4 and 8, the NAEP assessment results were mixed; therefore, there is no conclusive evidence that U.S. students are improving in mathematics—an important topic for further investigation.

Expressed regional need

State and local education stakeholders in the Mid-Atlantic region (Delaware, Maryland, New Jersey, Pennsylvania, Washington, DC) also identified improving mathematics achievement as a priority. They outlined a need for curricula and instructional practices aligned with their state standards for mathematics (National Center for Education Evaluation and Regional Assistance 2011). Based on these needs and priorities, Regional Educational Laboratory Mid-Atlantic reviewed four mathematics curriculum candidates that met this criterion and could be subjected to rigorous evaluation.

Connected Mathematics Project 2

The instructional intervention for this study is Connected Mathematics Project 2 (CMP2), developed over several years. The original CMP, developed over 1991–1996, was funded by the National Science Foundation to create a problem-centered mathematics curriculum designed for students in grades 6–8. According to the developers, the program was influenced by research in cognitive science, mathematics education, and educational policy and organization (Lappan et al. 2006a).

Between 2000 and 2005, the developers revised CMP, with the goals of increasing applicability and cohesion. To these ends, the developers examined the appropriateness of the language and readability of the materials for use with diverse student subpopulations, such as English language learners or students considered at risk of academic failure based on prior low achievement in mathematics (Lappan et al. 2006a). Each element of the curriculum was revised to improve cohesion of the units. Especially problematic units were dropped and replaced by new ones, which were then piloted and refined. The revised curriculum is CMP2.

CMP2 was selected for this study for three primary reasons: it aligned with expert opinions on best practices in mathematics instruction, it is widely used in the Mid-Atlantic region and across the United States, and only limited research evidence rigorously evaluated its effects, but enough to justify further research.

Aligned with National Council of Teachers of Mathematics standards

The National Council of Teachers of Mathematics Curriculum (NCTM) and Evaluation Standards of 2000 is one of the most influential sets of standards in mathematics education (Remillard, Herbel-Eisenmann, and Lloyd 2009; Schoen and Hirsch 2003). While not negating the importance of a procedural understanding of mathematics, the standards stress the

importance of conceptual understanding, connections to student experience and between mathematics topics, and an active role for students in their own learning.

Consistent with NCTM recommendations, mathematics educators and researchers have advocated curricula that engage students in mathematics problems that connect to the real world (Battista 1999; Boaler 2002; Elbers 2003; Martin 2007; Moss and Beatty 2006; Schoen and Hirsch 2003). Other researchers question such inquiry methods, arguing that they could place a high demand on working memory (Kirschner, Sweller, and Clark 2006). Direct instruction, for example, provides more explicit guidance, and some researchers advocate this method over inquiry methods (Przychodzin et al. 2004).

Quasi-experimental mathematics education research also supports practices consistent with NCTM recommendations. One important finding is that classroom discourse can allow for group reflection and sense making (Elbers 2003; Hodge et al. 2006; Moss and Beatty 2006; Ozmantar and Monaghan 2006; Powell 2006). Another is that students in traditional instruction curricula based largely on textbook series tend to compartmentalize mathematics topics, understand algorithms in rote ways, and separate mathematics from real-world contexts (Battista 2001; Cobb, Yackel, and Wood 1992; Lucas 2006; Popkewitz 2004).

Consistent with NCTM recommendations, the CMP2 developers designed the curriculum to focus on conceptual understanding, with topics connected to the real world and to each other (Lappan et al. 2006a). The implementation guide encourages teachers to support students in having an active role in their own learning and engaging in rich classroom discourse (Lappan et al. 2006a).

Widely used nationally and regionally

According to the publisher (M. Baughman, personal communication, January 2007), more than 3,500 U.S. school districts purchased CMP2 over 2005–2007.[15] About 700 districts were in the Mid-Atlantic region, representing 62 percent of all Mid-Atlantic districts.

Limited rigorous research evidence

In 2005, the What Works Clearinghouse (WWC)[16] reviewed research on CMP. Their report identified 22 studies that investigated its effects of CMP on middle school student outcomes. None of the studies “met WWC standards” (the designation for the most rigorous evaluations, which use randomized controlled trials (RCT) and have low attrition rates) and only three (Ridgway et al. 2003; Riordan and Noyce 2001; Schneider 2000) used quasi-experimental comparison-group designs, qualifying them to meet WWC evidence standards “with reservations” (the designation for studies with weaker evidence).

In January 2010, the WWC updated its review. [17] An additional 57 (for a total of 79) studies were identified as investigating the effects of CMP and CMP2 on middle school student outcomes.

[15] The publisher considers the exact number of school districts to be proprietary information, and it was not available to the authors.
[16] The What Works Clearinghouse was established in 2002 by the Institute of Education Sciences to provide educators, policymakers, researchers, and the public with a central and trusted source of scientific evidence of what works in education. http://ies.ed.gov/ncee/wwc/
[17] WWC did not differentiate between CMP and CMP2 in their 2010 review.

The updated review applied more stringent criteria for baseline equivalence between groups. In the new review, none of the studies met WWC evidence standards, and only Schneider (2000) met evidence standards "with reservations."

Following is a summary of the three quasi-experimental studies that originally met WWC evidence standards with reservations and the one experimental study that has been published since the current study began.

Schneider (2000)

Schneider (2000) compared the performance of students in grades 6–8 in 23 CMP schools (6,557 students) and 25 matched comparison schools (5,605 students) in Texas. CMP and comparison schools were matched on their predicted values on the 1996 Texas Assessment of Academic Skills. Predicted values for CMP and comparison schools were estimated from regression models with the following school-level covariates: percentage of students passing in mathematics on the 1995 and 1996 assessment, percentage of students present in school for less than 83 percent of the school year (counted as percentage of student mobility) in 1996, and total school enrollment in 1996. The schools were then sorted by their predicted values. A school was selected to be a control school if it had the closest predicted value to the predicted value of a CMP school and had the same grade levels. The study found that students attending CMP schools scored statistically significantly lower than the students in the matched comparison schools (effect size = –0.14; $p < .05$). However, the analysis was not conducted at the level of assignment (schools), resulting in a mismatch between the unit of assignment and unit of analysis, an underestimate of the standard errors (due to inflated samples size), and an overestimate of the statistical significance (or *p*-values). After statistically correcting for the mismatch, the WWC found no statistically significant differences between the two groups at $p < .05$.

Riordan and Noyce (2001)

Riordan and Noyce (2001) compared the mathematics achievement of grade 8 students in 20 CMP schools (1,879 students) and 30 matched comparison schools (4,978 students) in Massachusetts. CMP schools were defined as schools that had completed at least 11 units[18] of CMP in grades 6–8 by 1998/99, as determined through phone interviews. CMP and comparison schools were matched on average Massachusetts Educational Assessment Program scores and the percentage of students eligible for free or reduced-price lunch.

Researchers found that the CMP group outperformed the comparison group. While a reanalysis of the data by the WWC in the original review confirmed this positive effect of CMP (effect size = 0.43), it also revealed that it was not statistically significant at $p < .05$. This study failed to meet the WWC standards in the 2010 review because it did not demonstrate that CMP and comparison schools were equivalent at baseline in the analytic sample.

[18] A year's curriculum for a particular grade is organized into eight thematic units. The 11 units referred to in this study extend over multiple years of use. The structure of CMP2 is presented in the following section.

Ridgway et al. (2003)

Ridgway et al. (2003) compared middle school students' mathematics achievement in CMP and comparison schools for two school years. Schools were matched on location, student population diversity, and student ability. The study included students in grades 6 and 7 in 1994/95 and students in grade 8 in 1995/96. The 1994/95 sample included nine CMP schools (340 grade 6 students; 630 grade 7 students) and nine comparison schools (160 grade 6 students; 250 grade 7 students). The 1995/96 sample included an undisclosed number of CMP schools (grade 8 n = 780 students) and an undisclosed number of comparison schools (grade 8 n = 300 students). The grade 6 students were new to the program, and about three-fourths of the grade 7 and grade 8 students had used the program the previous year, as the study was conducted in middle schools currently implementing CMP. The Iowa Test of Basic Skills Survey Battery (ITBS) and the Balanced Assessment of Mathematics (BAM) were used to measure student performance in the fall and spring of each year.

The results were mixed. The grade 6 CMP group was 1.0 grade equivalence level behind the comparison group in the fall ($p < .001$) and 1.5 behind in the spring ($p < .001$). Using ANOVA, the interaction was statistically significant ($p < .01$), indicating that the gains were different. The grade 7 CMP group was 0.5 grade equivalence level behind the comparison group in the fall ($p < .001$) and 1.5 behind in the spring ($p < .001$). An ANOVA demonstrated that the interaction was not statistically significant and that the gains for each group were not different ($p < .28$). The grade 8 CMP group was 0.5 grade equivalence level ahead of the comparison group in the fall ($p < .001$) and 0.8 ahead in the spring ($p < .001$). Based on an ANOVA, the interaction was not statistically significant and there was thus no difference in the gains between the groups ($p < .053$). On the BAM, CMP students outperformed comparison students across all grade levels, with effect sizes of 0.15 for grade 6, 0.53 for grade 7, and 0.80 for grade 8. All differences were statistically significant at $p < .001$.However, according to the 2010 WWC review, Ridgway et al. had not demonstrated that CMP and comparison schools were equivalent at baseline in the analytic sample, so the study no longer met WWC standards.[19]

Eddy et al. (2008)

Eddy et al. (2008) was released after WWC compiled studies for the 2010 review and was not included in that summary. For this reason, and because at the time of this writing Eddy et al. (2008) was the only published RCT on CMP2, it is summarized here with slightly more depth than the WWC reviewed quasi-experimental studies.

Eddy et al. (2008) chose six middle schools across three states for which to compare the mathematics achievement of grade 6 students and their attitudes toward mathematics in 2007/08. At each school, teachers were randomly assigned to either a treatment or control group. Student mathematics achievement was measured by the BAM and the ITBS.[20] A student survey was used to measure student attitudes toward mathematics and to ask students about their teachers' characteristics.

---

[19] The 2010 WWC review only considered the grade 6 sample, because many of the grade 7 and grade 8 students had previous experience with CMP.

[20] The researchers report results for ITBS as grade equivalence scores and as norm curve equivalent scores. Results from the impact analyses did not differ by score type. Therefore, to simplify this discussion, score types are not differentiated when reporting results for the ITBS.

The study sample included 11 CMP teachers (509 students) and 9 control teachers (405 students). Eddy et al. (2008) reported overall student attrition (from pre- to posttest) at 18 percent and differential student attrition between treatment and control groups (from pre- to posttest) at 1 percent. Student overall attrition rates were 21 percent each for the ITBS, BAM, and student attitude survey; differential attrition rates were 5 percent for the ITBS and 1 percent for the BAM and student attitude survey. According to their implementation findings, 20 teachers participated the entire year of implementation.

Both repeated measures ANOVA and hierarchical linear modeling (HLM) were used to empirically assess the effect of CMP2 on student outcomes. However, the repeated measures ANOVA analysis was limited for several reasons. The analysis was conducted at the student level despite teachers being the level of assignment, did not correct the reported *p*-value for clustering, did not differentiate variance within teachers from variance between teachers, and did not control for student, teacher, or classroom characteristics. Eddy et al. (2008) acknowledge these limitations. Therefore, the results from the HLM analysis (with students at level 1 and classrooms at level 2) is the focus of this discussion and the basis for the conclusions.

The study did not detect an impact of the intervention either overall or for any of the examined student subgroups. The HLM analysis conducted for the ITBS, BAM, and the student attitude survey revealed no statistically significant differences between students in treatment and control conditions after controlling for student and teacher characteristics and accounting for the clustering of students within classrooms. Further, there were no statistically significant effects of CMP2 on outcomes for student subgroups based on gender or ethnicity or on teacher characteristics (specifically, number of years of teaching experience and teacher efficacy).

This study has three main methodological limitations. First, the HLM analyses were not specified fully to correspond with the within-school random assignment design. For example, all the HLM analyses used a two-level model without a third level to represent schools, which might result in underestimated standard errors and overstated *p*-values. However, it is unlikely to have changed the conclusions since the researchers reported statistically insignificant results. Second, the small number of teachers (n = 20) resulted in low statistical power requiring a large effect size to achieve statistical significance. Third, because teachers were randomly assigned within schools, CMP2 and control teachers were within the same school, resulting in possible diffusion of CMP2 instructional practices and curriculum use to control teachers and students. This could reduce the contrast between CMP2 teacher instructional practices and control teacher instructional practices, which in turn could reduce the magnitude of the impact estimate. Whether this reduction occurred cannot be determined because the report did not include an analysis for diffusion.

Summary of existing research evidence

The evidence on the effect of CMP and CMP2 on mathematics achievement is inconclusive. Of the three quasi-experimental studies reviewed here, two did not adequately equate groups at baseline, and the third shows no statistically significant differences between groups once appropriate corrections are made (What Works Clearinghouse 2010). The only experimental study of CMP2 also found no statistically significant differences (Eddy et al. 2008).

All the reviewed studies had methodological limitations. This study addresses these methodological limitations and extends the validity of previous research to the Mid-Atlantic region.

Description of CMP2

The CMP2 curriculum

The CMP2 curriculum has a problem-centered learning approach. The approach takes the form of investigations, in which mathematical concepts are embedded in problems thought to be interesting to the student, such as determining which basketball player is better at free throws, Yao Ming or Shaquille O'Neal.[21] A year's curriculum is organized into eight thematic units (table 1.1). Each unit has one to five investigations, each investigation has multiple lessons, and each lesson has three instructional phases (Launch, Explore, and Summarize). These components are described below.

Units

Grade 6 units are organized around core mathematical strands.[22] Some units culminate in an independent project. For example, at the beginning of the Prime Time unit, students complete an independent project called "My Special Number Project," in which they choose a number between 10 and 100 and write several things about the number. After each investigation, they apply the new concepts they have learned (such as prime or composite, factors, multiples, common factors, and common multiples) to write more about their special number.

The implementation guide recommends at least 50 minutes per day to complete the grade 6 curriculum (at least 6 of 8 units) in one school year.

Table 1.1. CMP2 grade 6 units

Unit title	Mathematics content	Recommended length of unit (days)
Prime Time	Factors and multiples	22.0
Bits and Pieces I	Understanding rational numbers	24.0
Shapes and Designs	Two-dimensional geometry	24.0
Bits and Pieces II	Understanding fraction operations	22.0
Covering and Surrounding	Two-dimensional measurement	27.0
Bits and Pieces III	Computing with decimals and percentages	29.0
How Likely Is It?	Probability	19.0
Data About Us	Statistics	19.5

Note: Length of unit in days assumes the recommended class period per day of approximately 50 minutes.
Source: Lappan et al. 2006a,b.

[21] "Who's the Best" investigation 4.1 in Bits and Pieces I introduces students to percentages as a part-whole relationship and uses students' previous experience with fraction partitioning and benchmark fractions to make sense of percentages.
[22] Trainers provided by the publisher work with schools to target the units that need to be finished during the year to meet grade-level standards.

Investigations

Each unit has several investigations based on challenges one might encounter in life. Students learn important mathematics concepts and procedures from these investigations. Each investigation takes a few days.[23] An example of a challenging problem that a teacher would present for the students to investigate is as follows:

> Jeremy and his little sister Deborah are at a carnival, and each rides a different sized Ferris wheel. One Ferris wheel is rotating every 20 seconds, and the other one is rotating every 60 seconds. Jeremy and Deborah take off simultaneously on their respective rides from the same initial starting position. Students then investigate how long it will take until both of the children once again concurrently reach the initial point at the bottom of the ride. Students explore finding common multiples of 20 and 60 to solve the problem.[24]

Lessons

The three phases of each lesson—Launch, Explore, and Summarize—provide structure to each lesson.

- In *Launch*, the teacher introduces the challenge problem, defines the mathematical goals, and connects new content to prior understanding and investigations.
- In *Explore*, the students share ideas with each other, justify their thinking, and develop solution strategies together. The teacher circulates around the class to ask questions, provides clarification and redirection, and assist students when needed. The teacher does not provide solutions.
- In *Summarize*, students present their solution strategies to the class and justify their thinking. The teacher connects the different presentations to prior learning and core mathematics concepts (appendix A).

Additional problem sets—called applications, connections, and extensions—are provided in each investigation to help students practice their understanding and skills at the discretion of the teacher and can be used as homework at the end of an investigation. The applications provide students with a short problem situation followed by a question. Students answer the question and explain their reasoning. Connections are a series of problems that allow students to connect new learning to prior concepts and skills (such as practice with operations on fractions) in both open-ended and multiple-choice format. Extensions are challenge problems for students to apply what they have learned to novel problem situations.

---

[23] For example, the Prime Time unit has five investigations, with each investigation lasting 1.5 to 3 days.

[24] “Riding Ferris Wheels,” investigation 3.1 of Prime Time, has students explore situations where finding common multiples of whole numbers is important.

Teacher resources typically available with CMP2

Professional development

The publisher offers several options for professional development (PD). One of the standard packages includes five days of PD: two during the summer before using the curriculum and three during the school year. This package was selected for the current study because the publisher considers it consistent with a “typical” implementation of CMP2 (M. Baughman, personal communication, January 2007).

The publisher’s trainers facilitate all PD sessions. The two summer days cover the curriculum components and implementation guide, the overall goals of the curriculum, and the content and pedagogy for the initial units to be taught. During the year, the trainer organizes each day of PD to coincide with the beginning of a new unit. Each day of PD begins with a debriefing on completed units and continues with preparation for the upcoming unit. The trainer and teachers analyze the effectiveness of prior implementation and plan for changes to instruction, discuss the content of upcoming units, and plan for sequencing and emphasis to fit teachers’ needs and district schedules (for example, preparation for state testing).

Before the start of subsequent school years, the publisher offers all schools a two-day summer PD session for new teachers. If the school has no teachers with CMP2 experience, they also provide three sessions during the school year. If the school has teachers with CMP2 experience, they are expected to mentor the new teachers in lieu of these sessions. See appendix A for additional information on PD for CMP2.

Implementation guide

The implementation guide provides an overview of the program, suggestions for preparing for implementation, pacing guides for both standard (45–60 minutes) and block (90 minutes) period scheduling, recommendations for classroom management and addressing special needs of students, and strategies for involving parents and the community.

Teacher guides

Teacher guides for each unit include discussions of the mathematics of the unit and descriptions of each phase of the instructional model for each problem (including sample questions), materials needed, instructional strategies, connections to other units, technology, and assessment resources.

Online resources

Descriptions of the curriculum and additional materials, not expressly for teacher use, are available online.

CMP2 theory of change

The theory of change, based on the developers’ statements about CMP2 in the implementation guide (Lappan et al. 2006a), is illustrated in figure 1.1. Lappan et al. (2006a) state that they designed the curriculum to focus on conceptual understanding, with topics connected to the real world and to each other. The implementation guide encourages teachers to support students in

having an active role in their own learning and engaging in rich classroom discourse. The types of practices promoted by the developers might also improve student engagement. Research over the last two decades has shown a relationship between achievement and the following school-related variables: academic engagement, perceptions and attitudes, self-confidence in learning mathematics and science, interest and motivation to learn mathematics and science, and self-efficacy (Eccles and Jacobs 1986; Helmke 1989; Reynolds and Walberg 1991, 1992).

Instructional practices taught in the CMP2 PD are expected to change teachers' classroom practices from an approach that focuses on procedure learning and practice to one that encourages connections between procedures and a real-world problem and emphasizes student-led discussions and problem-solving. This should change how students learn mathematics, as they learn to communicate with each other effectively while developing mathematics knowledge. Finally, these classroom-level changes are expected to improve mathematics achievement and the value that students place on mathematics. This study focuses solely on the impact of CMP2 on these final outcomes.

Figure 1.1. Theory of change for CMP2

Teacher PD →	**Teacher classroom practices** →	**Intermediate outcomes** →	**Final outcomes**
Teachers learn: • Why changing practice could help students learn. • How to implement a balanced, real-world curriculum. • Strategies for classroom discussion. • The CMP2 curriculum, including lessons and homework assignments.	**Teacher classroom practices include:** • Connecting "real world" problem to mathematics concept or procedure. • Facilitating student-led discussions and problem-solving. • Encouraging different approaches to solving problems	**Student practices include:** • Increased practice with using standard mathematics tools in the context of real-world problems. • Increased work in difficult content areas (for example, probability). • Increased practice in communicating effectively.	**Improved student mathematics achievement** **Improved value that students place on mathematics**

Source: Authors' interpretation of the theory of change from Lappan et al. (2006a).

Defining the counterfactual: Business as usual

Since control schools in this study were expected to continue business as usual; that is, to continue using their respective mathematics curricula, it was important to define the expectation for mathematics instruction in control classrooms and to identify how that activity would vary from the mathematics instruction in intervention classrooms.

The most recent TIMSS videotape classroom study (Stigler et al. 1999) found that 90 percent of class time in observed U.S. mathematics classes was spent practicing routine procedures. The

predominant observed lesson model included a teacher demonstration of how to solve a procedural problem followed by student application of the procedure by solving examples independently. During the independent practice, many teachers helped students having difficulty. This model contrasted sharply with one in Japan, for example, where problem solving came first, followed by students reflecting on the problem, sharing the solution methods they generated, and jointly working to develop explicit understandings of the underlying mathematical concepts.

While it was expected that control classrooms would use a variety of mathematics curricula, the hypothesis was that these classrooms would be more likely to exhibit characteristics similar to those observed in the TIMSS videotape classroom study (table 1.2).

Table 1.2. Instructional activities hypothesized to be present in intervention and control classrooms

Activity	CMP2-like instruction hypothesized to be prevalent in intervention classrooms	Less CMP2-like instruction hypothesized to be prevalent in control classrooms
Making connections	The teacher makes connections between a real-world context or problem and the mathematics concept or procedure.	The teacher makes connections to rules and procedures learned previously, but no real-world or practical connection is attempted.
Teacher factors	Classroom seating is conducive to collaborative learning. The teacher asks questions rather than telling students what rules and procedures to apply. The teacher expects students to ask and answer each other's questions. The teacher encourages different approaches to solving problems.	Classroom seating is in rows. The teacher leads instruction, demonstrating the rules or procedures for the day. When students don't understand, the teacher shows or tells them the rules or procedures.
Student factors	Students talk to each other and share their thinking. It is common for students to introduce different ways to solve a problem and to ask each other questions about the different methods.	Students primarily work independently, seeking assistance from the teacher rather than talking to other students. The focus is on accuracy or correctness of work.
Emphasis	Majority of time spent on class discussions, small group, or pair work.	The majority of time is spent on teacher-led lecture and independent student work.

Source: Classroom observation protocols.

Study description

The current study is a multisite cluster RCT examining the impact of CMP2 on mathematics achievement and perceived mathematics task value scores of grade 6 students in Mid-Atlantic region schools. Although CMP2 is designed for students in grades 6 through 8, this study focuses on students in grade 6 because it represents the start of the curriculum series and is thus the most likely grade level of initial adoption. The study was implemented in 65 volunteer schools across the Mid-Atlantic region over two years. During the implementation year (2008/09), teachers received PD and gained experience using the CMP2 curriculum. During the impact year (2009/10), the intervention schools were expected to have fully implemented CMP2, along with

required teacher preparation. The study team collected student outcome data in the fall and spring of this year in both intervention and control schools.

Random assignment was conducted at the school level—with participating schools assigned to either the intervention group (CMP2) or the control group (business as usual)—because curricular implementation is typically at this level. This level of assignment also reduces the risk of sample contamination, as all teachers in a particular school will be in the same study group.

To increase statistical power, student pretest scores and other variables found to be statistically significant at baseline were included in the analytic model as covariates (see chapter 2).

This study also used classroom observations to determine whether the intervention was being implemented in the intervention schools and what differences, if any, could be observed between the intervention and control schools during the impact year (see chapter 3).

Research questions

This study was designed to address one primary and one secondary research question. The primary research question was:

What is the impact of being in a school randomly assigned to adopt CMP2 on grade 6 student mathematics achievement?

The secondary research question is related to the effect of CMP2 on students' perceived task value (PTV) of mathematics. It is a secondary question because, without statistically significant findings on a direct measure of mathematics achievement (the primary research question), the answer to the secondary research question alone would be insufficient for assessing the impact of CMP2. The secondary research question is:

- What is the impact of being in a school randomly assigned to adopt CMP2 on the value that grade 6 students place on mathematics?

Guide to subsequent chapters

The rest of the report is organized in four chapters. Chapter 2 describes the study design and methodology. Chapter 3 summarizes the data collected on fidelity of implementation, the analysis of the level of implementation in the intervention schools, and a comparison of the instruction in control and intervention schools. Chapter 4 compares the baseline characteristics of the control and intervention schools and provides the results from the main analyses. Chapter 5 summarizes and discusses the findings.

2. Study Design and Methodology

This chapter describes the study design and methodology, including sample recruitment, random assignment and changes in the study sample, attrition and characteristics of the study sample, data collection instruments, and data collection and analysis methods.

Study design

This study was conducted in 53 school districts in four of the five jurisdictions in the Mid-Atlantic region (Delaware, Maryland, New Jersey, Pennsylvania and Washington, DC) over 2008/09–2009/10. It used a cluster RCT design, in which 70 volunteer middle schools were randomly assigned either to use the CMP2 curriculum (intervention condition) or to continue with their existing approach to teaching mathematics (control condition). Schools were selected as the unit of random assignment because CMP2 encourages within-school communication and collaboration between teachers. Within school randomization of teachers or classrooms would have increased the likelihood of cross-group contamination. The random assignment of schools was performed separately for each jurisdiction.

The primary outcome was mathematics achievement as measured by the TerraNova CAT™2 Basic Multiple Assessments Form C (CTB/McGraw-Hill 2003; TerraNova). The students' perceived task value (PTV) of mathematics was measured by the PTV portion of an instrument designed by Eccles and Wigfield (1995). These measures were also used to evaluate student mathematics achievement and the value students place on mathematics at baseline.

Effectiveness trials evaluate the effect of interventions under typical, rather than optimal, conditions (Flay 1986; Flay et al. 2005). Consequently, this study was designed to take place in the instructional environment that would have occurred had school districts purchased and implemented CMP2 on their own. The study team monitored but did not interfere in the natural implementation of the program.

Study timeline

Recruitment began in January 2008. The study was divided into an implementation year (2008/09) and an impact year (2009/10; tables 2.1 and 2.2).

Implementation year

Use of an implementation year (2008/09) was motivated by research suggesting that changes in complex teaching practices, such as moving from direct instruction to the problem-centered curriculum of CMP2, requires approximately 25 practice sessions before teachers can comfortably integrate new techniques into their pedagogy (Joyce and Showers 1995).

During the implementation year, teachers in intervention schools were expected to participate in CMP2 PD and gain experience using the curriculum. The study team collected teacher- and classroom-level data through classroom observations and teacher monthly online surveys. Control schools continued business as usual and did not participate in data collection during this year. The data collection instruments and the data collection and analysis process are described in more detail at the end of this chapter.

Table 2.1. Timeline for key research activities preceding and during the implementation year (2008/09)

Date	Activity
January–April 2008	District and school leaders submitted letter of interest
April–May 2008	District and school leaders signed memorandums of understanding. Schools were randomly assigned to study conditions.
July–August 2008	Summer PD was conducted for teachers at intervention schools using a large group, 2-day session. Teacher consent forms and demographics were collected.
September 2008	Instruction began in intervention schools.
October 2008	Teachers at intervention schools completed first monthly online survey.
November 2008	Classroom observations were conducted at intervention schools. PD follow-up was conducted for teachers at intervention schools using a large group, 1-day session.
February 2009	Second PD follow-up was conducted for teachers at intervention schools using a large group, 1-day session.
April 2009	Second classroom observations were conducted at intervention schools. Third PD follow-up was conducted for teachers at intervention schools using a large group, 1-day session.
May 2009	Teachers at intervention schools completed last monthly online survey.

Source: Study records.

Impact year

The impact of CMP2 was estimated during the impact year (2009/10), when lab personnel[25] observed lessons in both intervention and control schools. Teachers in the intervention schools completed monthly online surveys, as well as a cumulative survey at the end of the year. Student data were collected from both intervention and control schools in the fall and spring of the impact year.

[25] Lab personnel were experienced education professionals who supported the study team by conducting observations and carrying out other tasks as outlined in this report.

Table 2.2. Timeline for key research activities during the impact year of the CMP2 effectiveness study (2009/10)

Date	Activity
June–September 2009	Class section rosters were collected to prepare for pretesting. New teachers in intervention schools were invited to summer PD using a large group, two-day session.
September 2009	Teacher consent forms and demographics were collected. Instruction began in all schools. Teachers at intervention schools completed first monthly online survey.
September–October 2009	Student assent and parent passive consent procedures were completed. Pretest data were collected on students in intervention and control schools.
November 2009	Classroom observations were conducted in intervention and control schools. First PD follow-up was conducted for new teachers at intervention schools.[a]
January 2010	Second PD follow-up was conducted for new teachers at intervention schools.[a]
February 2010	Third PD follow-up was conducted for new teachers at intervention schools.[a]
February–March 2010	Class section rosters were collected to prepare for posttesting. Classroom observations were conducted for intervention and control schools.
May–June 2010	Teachers at intervention schools completed last monthly online survey. Posttest data were collected on students in control and intervention schools.

a. PD during the school year for new teachers was contingent on the unavailability of an experienced CMP teacher at the school to act as a mentor.
Source: Study records.

Target population and sample recruitment

A statistical power analysis indicated that a minimum of 67 schools and 4,623 students[26] were required for a minimum detectable effect size (MDES) of 0.20 standard deviations. This MDES was considered reasonable based on the CMP research available when this study was designed. The average of the absolute values of the effect sizes reported in Schneider (2000), Riordan and Noyce (2001), and Ridgeway et al. (2003), was 0.24 standard deviations (see appendix B for more information on this analysis). To buffer against potential attrition, the study team planned to recruit 70 schools.

School eligibility criteria

The study was open to all public and charter schools in the Mid-Atlantic region that met the following eligibility criteria:

- The school was not using, nor had used, CMP or CMP2.
- The school agreed to be randomly assigned to either the intervention or control condition.

[26] This number was calculated based on estimates of 3 class sections per school grade and 23 students per class section (69 students per school).

- The school agreed to participate in study activities for two years.
- If assigned to the intervention condition, the school agreed to participate in the PD provided by the publisher.
- The school agreed to allow only general education grade 6 mathematics teachers and class sections to participate in research activities. Special education students and English language learner students, for example, would be included in the study only if they were enrolled in general education grade 6 mathematics classes.
- The school agreed to adhere to the study protocols for key research activities, such as student testing and classroom observations.

Schools assigned to the intervention group received free CMP2 curriculum materials and free PD for both study years, and each intervention teacher received $25 at the end of each year for completing the survey. Control schools received $1,000 each at the completion of the study.

Recruiting to achieve the target analytic sample size

Recruitment began with building awareness in the region through forums and presentations and continued with contacting targeted schools through letters, phone calls, and meetings. This section describes each step in recruitment, from identifying potential participating schools through receipt of letters of interest to participate in the study, to obtaining a signed memorandum of understanding in preparation for randomization.

Identifying schools

A list of schools serving grade 6 students in the Mid-Atlantic region was derived from the Common Core of Data (CCD; National Center for Education Statistics n.d.a) in August 2007. Schools believed to be using CMP2 were removed from the list, based on data provided by the CMP2 publisher.[27] In January 2008, the study team sent invitations to 989 potentially eligible districts with 2,597 schools serving grade 6 students (table 2.3). Lab personnel and principal investigators either called or met with superintendents, principals, teachers, and other school staff to explain the purpose of the study, the CMP2 curriculum, and schools' required time investments for participation.

Letter of interest

If a district or school was interested in participating in the RCT, the superintendent or a designee signed a letter of interest containing the name of the school, the number of grade 6 teachers, and the number of class sections available. Letters of interest were received from 105 schools across a total of 80 districts in the Mid-Atlantic region.

District and school administrators were contacted by phone or e-mail to further clarify the study requirements, eligibility criteria, and incentives and to determine if the school was qualified and interested in signing a memorandum of understanding before being selected to participate in the study. After the contacts, 32 of the 105 schools withdrew their letter of interest because they

[27] The publisher provided general data on purchase history but could not confirm usage. More specific information was considered proprietary and thus not included in the report.

were unable to meet all the participation requirements. The 73 remaining schools moved to the memorandum of understanding phase of recruitment.

Memorandum of understanding

The memorandum of understanding was an agreement to participate in the study, including all data collection activities. It described the eligibility criteria and conditions for participation. It explained that if assigned to the intervention group, participating schools would have access to the CMP2 curriculum for grade 6 and that if assigned to the control group, participating schools would not have access.

Seventy-three schools were invited to sign the memorandum in March 2008, and 72 did. The number fell to 70 before randomization due to the merger and withdrawal of schools.[28]

Table 2.3. Sample sizes at different stages of recruitment

Recruitment activity	Number of districts	Number of schools	Percentage of school sample (n = 2,597)	Percentage of schools retained from previous recruitment activity[a]
Invitations mailed	989	2,597	100	na
Number contacted with follow-up calls, e-mails, faxes or meetings	853	470	18	18
Number that submitted a letter of interest	80	105	4	22
Number that did not withdraw letter of interest	54	73	3	70
Number signed memorandum of understanding	53	72	3	99
Number in random assignment pool	53	70[b]	3	99

na is not applicable
a. The number of schools in the current “recruitment activity” row divided by the number of schools in the previous “recruitment activity” row. For example, in the “number that submitted a letter of interest” activity row, this percentage is 22 ([“Number that submitted a letter of interest” = 105/ “Number contacted with follow-up calls, emails, faxes or meetings” = 470] x 100 = 22 percent).
b. The reduction in the number of schools at random assignment was due to either a school merger or a withdrawal from the study.
Source: Study records.

[28] Specific information on these participants’ administrative actions cannot be included in this report because it could compromise the confidentiality of the study participants and disclose a participating school’s identity.

Randomization of schools

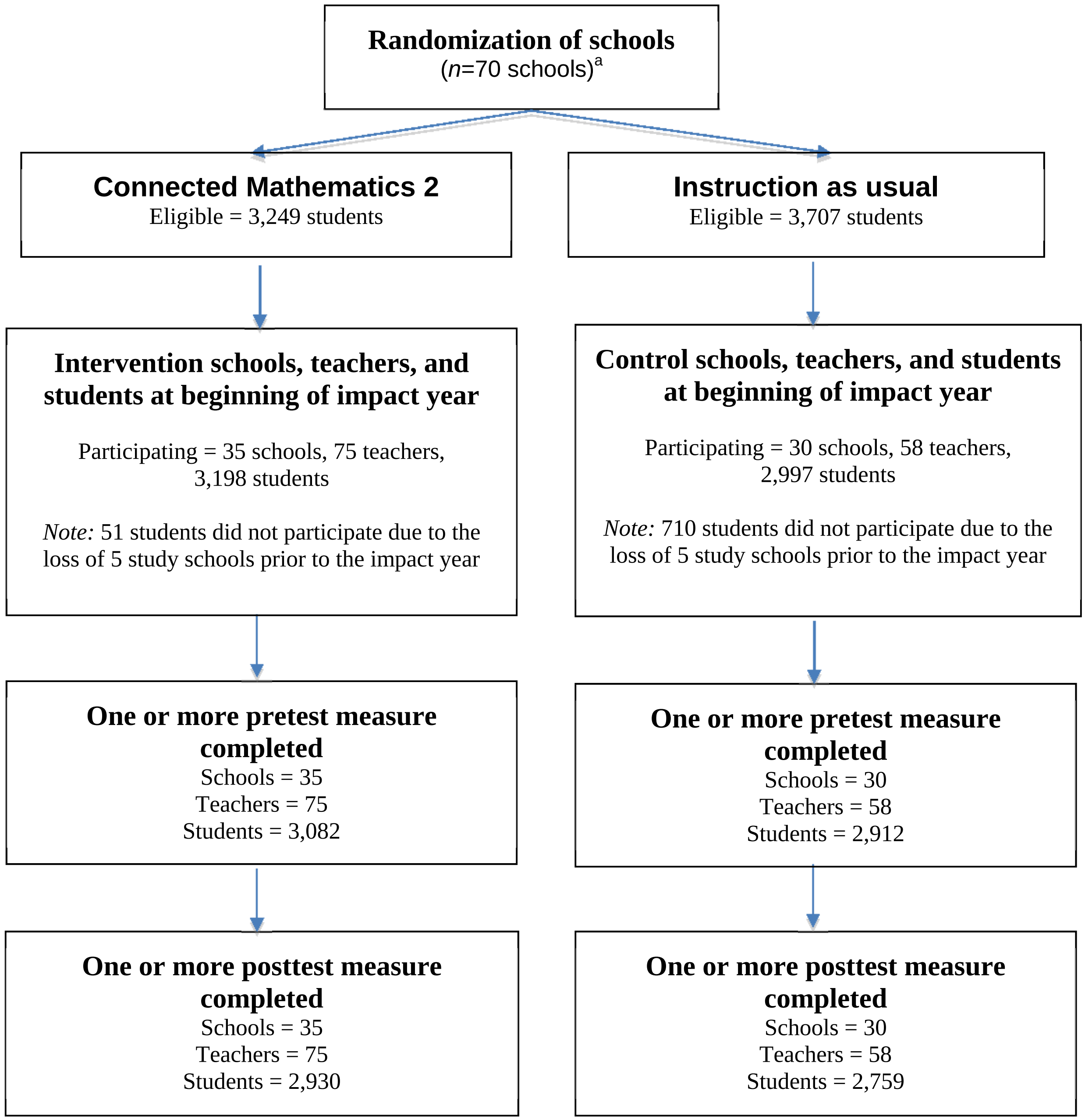
On average, random assignment produces two groups that are similar on both observed and unobserved characteristics.[29] While this is true at the time of randomization, group characteristics may change over time due to loss of participants (attrition) or other unexpected events. If these events are random, the integrity of the sample will be maintained. If not, the results of the study might be affected. For this reason, it is important to monitor the integrity of random assignment. Throughout the study, the numbers of participating schools and students were carefully tracked. Seventy schools were randomly assigned within jurisdiction to study conditions in May 2008. The imbalance between groups was the result of chance (figure 2.1).

Sample changes following randomization

Between the point of random assignment and the start of the impact year (2009/10), five schools were lost due to either a school-level administrator decision to withdraw the school or a district-level decision to close and merge campuses with low enrollment.[30]

[29] Randomization balances groups on measured and unmeasured characteristics over many repeated samples, but there can be statistically significant differences in any one particular sample, by chance (Friedman, Furberg, and DeMets 1997; Piantadosi 2005; Shadish, Cook, and Campbell 2002).
[30] Additional information cannot be provided on the specific reason for the loss of schools due to a potential disclosure of the identity of the schools.

Figure 2.1. Sample size at various stages of the study



a. The analytic sample, which includes students, teachers, and schools, was not established until the impact year/pretest (2009/10). Schools were the level of randomization and analysis. Before the impact year, five schools were lost due to withdrawal or merger with an existing study school.
Note: The analytic sample consisted of students who were pretest-eligible and had completed at least one posttest. Missing pretest data were imputed using the dummy variable adjustment approach. The analytic sample varied between the TerraNova and PTV measures because they were not administered on the same day at all locations, resulting in uneven participation numbers. Student numbers reported here represent those who completed at least one measure, either TerraNova, PTV, or both.
Source: Study records.

Changes between the implementation year and the impact year/pretest

At the beginning of the impact year, 65 schools (93 percent of the randomized sample) and 6,195 students remained in the sample, including 35 intervention schools (97 percent) with 3,198 students and 30 control schools (88 percent) with 2,997 students. No schools were lost during the impact year. Although the number of schools in the analytic sample was less than the 67 called for in the statistical power analysis, the final number of students (5,689) was 1,066 more than the target.

Before pretesting, seven intervention teachers chose not to implement CMP2. These teachers were permitted to proceed with business as usual. Due to the intent-to-treat analytic approach used in this study, these teachers were retained in the intervention group for analysis. These teachers agreed to have their students participate in pre- and posttesting, and their students' data were analyzed as intervention students in the analysis. These teachers did not participate in classroom observations, the monthly survey, or the end-of-year survey—and those data are considered missing. They did complete a demographic survey, and those data were retained in the intervention group for analysis.

Student participation

This section presents information on student participation. It tracks the student sample from pretest through posttest, the length of the impact year. No student data were collected during the implementation year. Specific data related to student participation rates are also included.

Student eligibility was determined at three stages of the study: baseline, pretest, and posttest.

- To be considered baseline eligible, a student had to be present in a study school at the beginning of the impact year before the five schools were lost.
- To be pretest eligible, a student had to be enrolled in a general education grade 6 mathematics class in a study school, agree to testing, and obtain parental consent.
- To be posttest eligible, a student had to have been pretest eligible (whether or not he or she actually participated in the pretest), still be enrolled in a general education grade 6 mathematics class in a study school, agree to testing, and obtain parental consent.

Student sample at pretest

Of the 70 eligible schools and 6,956 baseline-eligible students (3,249 intervention and 3,707 control), 65 schools and 6,168 students (3,184 intervention and 2,984 control) were included in the pretest sample, and 5,994 of the pretest sample completed at least one pretest measure (table 2.4).[31]

[31] The reduction in number of students from "baseline eligible" to "participating" was caused by the removal of 761 students due to the withdrawal or merger of five schools before the impact year and the loss of 27 students (14 intervention and 13 control) because of lack of student assent or parent consent.

Table 2.4. Summary of student participation in pretest data collection

Assessment	Pretest eligible intervention students (n = 3,184)		Pretest eligible control students (n = 2,984)	
	Pretested	Missing (due to absence)	Pretested	Missing (due to absence)
Both TerraNova and PTV	2,693	102	2,645	72
TerraNova only	316	175	220	118
PTV only	73	418	47	291
Total[a]	3,082	102	2,912	72

a. The sum of the number of students pretested and the number of students missing due to absence equals the *n* for each group.
Source: Study records.

Changes between pretest and posttest

Loss of students. Between the pretest and posttest, 403 students (224 intervention and 179 control) withdrew. Because students were required to have been both pretest eligible and to have completed at least one posttest measure to be included in the analytic sample,[32] these 403 students were not included in the analytic sample and were counted as attrition at the student level.

Ineligible students. There were 76 students identified as no longer in regular grade 6 mathematics classes at posttest (30 intervention and 46 control), so they were determined ineligible for the study. These students were not included in the analytic sample for this study and are not counted as attrition.

Student sample at posttest

The students included in the posttest administration were pretest eligible and remained eligible at posttest (table 2.5). Sixty-five schools (5,689 students) participated in the posttest data collection.

Make-up test opportunities were provided to students absent during posttest data collection. Multiple visits were made to the schools so absent students could complete the TerraNova test or PTV posttest survey, as needed. After repeated attempts to gather all possible posttest data, 12 students (7 intervention and 5 control) had still not completed the TerraNova posttest, and 105 students (32 intervention and 73 control) had not completed the PTV posttest survey.

[32] A student who was pretest eligible but did not participate would still be included in the analytic sample if he or she completed at least one posttest measure.

Table 2.5. Summary of student participation in posttest data collection

Assessment	Intervention students able to be tested (*n* = 2,930)		Control students able to be tested (*n* = 2,759)	
	Posttested	Missing (due to absence)	Posttested	Missing (due to absence)
Both TerraNova and PTV	2,891	0	2,681	0
TerraNova only	32	7	73	5
PTV only	7	32	5	73
Total[a]	2,930	0	2,759	0

Note: n = number of posttest eligible students (students in a general education grade-6 mathematics classroom in one of the 65 participating schools who assented to participate, had parental consent for the study, and had been pretest eligible.
a. The sum of the number of students posttested and the number of students missing due to absence equals the *n* for each group.
Source: Study records.

Response rates, attrition, and baseline equivalence

This section presents data on the analytic sample, including student response rates at pretest and posttest and attrition data. It then discusses baseline school characteristics and equivalence over time, first comparing the population with the randomized sample and then comparing the randomized sample to the analytic sample. The intervention and control schools are compared at baseline, and covariates to be applied to the statistical model are identified.

Analytic sample. The analytic sample included any pretest-eligible students (whether or not they participated) who remained eligible at posttest and completed at least one posttest measure (table 2.6). Any students in the analytic sample missing one or both pretest measures were retained in the sample and adjusted for in the analysis using the dummy variable adjustment approach (Puma et al. 2009). [33]

[33] A missing pretest score was replaced by a constant in the analytic sample and denoted in the multilevel model with a dummy variable coded as 1 to indicate that the actual pretest score was missing.

Table 2.6. Number of students with valid pretest and posttest scores, by assessment

	Posttest		
Pretest	**Either TerraNova or PTV**	**Both TerraNova and PTV**	**Total**
TerraNova only	10	448	458
PTV only	7	101	108
Both	96	4,932	5,028
Neither	4	91	95
Total	117	5,572	5,689

Note: For posttest, there is no "Neither" column, because students who did not complete at least one posttest were excluded from the analytic sample.
Source: Study records.

Response rates for each student data collection activity during the impact year ranged from 87 percent to 100 percent (table 2.7).

Table 2.7. Student response rates from pretest to posttest for students for the impact year

	Total[a]		**Intervention**		**Control**	
Status	**Number**	**Percentage of eligible**	**Number**	**Percentage of eligible**	**Number**	**Percentage of eligible**
Baseline eligible sample	6,956	100.00	3,249	100.00	3,707	100.00
Remaining in sample after loss of five schools	6,195	89.06	3,198	98.43	2,997	80.85
Able to be tested (consented)						
Pretest	6,168	88.67	3,184	98.00	2,984	80.50
Posttest	5,689	81.79	2,930	90.18	2,759	74.43
TerraNova						
Pretest	5,874	84.45	3,009	92.61	2,865	77.29
Posttest	5,677	81.61	2,923	89.97	2,754	74.29
Both	5,475	78.71	2,803	86.27	2,672	72.08
PTV						
Pretest	5,458	78.46	2,766	85.13	2,692	72.62
Posttest	5,584	80.28	2,898	89.20	2,686	72.46
Both	5,043	72.50	2,563	78.89	2,480	66.90

a. Baseline-eligible students are defined as those present in study schools at the beginning of the impact year prior to the loss of five schools due to withdrawal from the study or merger with an existing study school.
Note: Response rates are calculated as percentages, with the number of baseline-eligible students as the denominator.
Source: Study records.

Attrition rates from random assignment to analysis

The methodological literature does not provide empirical guidance on when exceeding thresholds for missing data introduces bias. Therefore, attrition rates are reported by intervention and control groups at the school and student levels; results for tests of statistical significance in these differences are reported to allow the reader to judge whether there is cause for concern about bias in the impact estimate. The attrition rates for intervention and control schools are provided in table 2.8. For example, 70 schools were randomized to the intervention and control conditions. Of those schools, 65 participated in the pretest and posttest. Therefore, the attrition rate from randomization to posttesting was 7.1%. The attrition rates for students for each assessment are provided in table 2.9.

Table 2.8. Number of schools and attrition rates for intervention and control groups for TerraNova and PTV

	Schools			
Data collection sample	**Intervention**	**Control**	**Difference**	**Total**
Schools in random assignment pool	36	34	2	70
Schools where pretests were conducted	35	30	5	65
Schools where posttests were conducted	35	30	5	65
Attrition from randomization to posttest (percent of schools)	2.8%	11.8%	9%[a]	7.1%

a. Based on the results of a *z*-test of proportions, the proportion (2.8 percent or 0.028) of school-level attrition for intervention schools was not statistically significantly different than the proportion of school-level attrition for control schools (11.8 percent or 0.118; $z = -1.46$, $p = .145$).
Source: Study records.

Because schools were lost, attrition rates differed in intervention and control schools. The overall attrition rate across intervention and control schools was 7 percent. The differential attrition rate between intervention and control schools was 9 percent. School-level attrition matters more than student-level attrition, because schools were the unit of random assignment and the impact analysis measures the impact of CMP2 at the school level. A statistically significant difference in the attrition rates between intervention and control schools would cause concern about potential bias in the impact estimate. However, this difference was not statistically significantly different ($p = .145$).

The student-level attrition rate for the TerraNova was 10 percent for the intervention group and 26 percent for the control group (a differential attrition rate of 16 percent; table 2.9). Similarly, for the PTV outcome measure, attrition was 11 percent for the intervention group and 28 percent for the control group (a differential attrition rate of 17 percent). The overall attrition rates were 18 percent for TerraNova and 20 percent for PTV.

This differential attrition rate is explained primarily by the withdrawal or merger of five schools from the study before the impact year began.

Table 2.9. Number of students and attrition rates for intervention and control groups for TerraNova and PTV

Data collection sample	Students							
	TerraNova				PTV			
	Intervention	Control	Difference	Total	Intervention	Control	Difference	Total
Baseline eligible students[a]	3,249	3,707	–458	6,956	3,249	3,707	–458	6,956
Students not tested due to withdrawal or merger of 5 schools	51	710	–659	761	51	710	–659	761
Students who refused assent/consent	14	13	1	27	14	13	1	27
Students with assent/consent	3,184	2,984	200	6,168	3,184	2,984	200	6,168
Pretested students	3,009	2,865	144	5,874	2,766	2,692	74	5,458
Posttested students	2,923	2,754	169	5,677	2,898	2,686	212	5,584
Student attrition from eligible to posttest	326	953	–627	1,279	351	1,021	–670	1,372
Student attrition rate from eligible to posttest (percent)	(10.0)	(25.7)	(–15.7)[b]	(18.4)	(10.8)	(27.5)	(–16.7)[c]	(19.7)

a. “Baseline eligible” students are those in study schools at the beginning of the impact year before the loss of five schools due to withdrawal from the study or merger with an existing study school.
b. Based on the results of a z-test of proportions, the proportion (10.0 percent or 0.100) of student-level attrition for intervention students was statistically significantly different ($z = –17.64$, $p < .001$) than the proportion of student-level attrition for control students (25.7 percent or 0.257) for the TerraNova.
c. Based on the results of a z-test of proportions, the proportion (9.4 percent or 0.094) of student-level attrition for intervention students was statistically significantly different ($z = –20.18$, $p < .001$) than the proportion of student-level attrition for control students (27.5 percent or 0.275) for the PTV survey.
Source: Study records.

Baseline school characteristics and equivalence over time

To investigate whether the 70 schools that volunteered to participate in the study were different from a larger population of eligible schools, the two groups were compared on characteristics hypothesized to be at least moderately correlated with students’ mathematics achievement, using the 2008/09 CCD (NCES n.d.b). Results for the sample of potentially eligible schools (n = 2,597) and the sample at random assignment (n = 70) indicate no statistically significant differences on examined school characteristics (table 2.10).

Table 2.10. Baseline school characteristics from recruitment to random assignment

School characteristic	Eligible schools (*n* = 2,581)[a]	Random assignment (*n* = 70)	Difference (SE)	Test statistic[b]	*p*-value
*Percentage of schools in each locale (*n *= 2,559)*[c]				$\chi^2 = 2.25$[d]	.523
Urban	24.23	24.29	–0.06	0.06[e]	.954
Suburban	49.36	42.86	6.50	1.31[e]	.190
Rural	18.91	25.71	–6.80	0.99[e]	.321
Small city	7.50	7.14	0.36	0.10[e]	.924
*School average student characteristics (*n *= 2,484)*[f]					
Average number of students enrolled	531.83 (322.35)	515.86 (277.92)	15.97 (33.84)	0.41	.682
Background characteristics of enrolled students					
Female (percent)	48.68 (3.84)	49.01 (3.39)	–0.33 (0.41)	–0.71	.480
Racial/ethnic composition (percent)					
White	54.81 (38.31)	58.33 (36.76)	–3.52 (4.46)	–0.76	.448
Black	28.34 (34.21)	25.69 (30.32)	2.64 (3.69)	0.64	.523
Hispanic	12.58 (19.72)	12.69 (12.95)	–0.11 (1.60)	–0.07[g]	.943
Asian or American Indian[h]	4.20 (6.93)	3.17 (5.39)	1.03 (0.66)	1.24	.216
Eligible for free or reduced-price lunch (percent)	33.88 (27.90)	34.78 (29.12)	–0.90 (3.53)	–0.27	.790

a. Of the 2,597 potentially eligible schools, school characteristic data from the CCD (2008/09) were available for 2,581 schools, as some schools had opened recently and their data were not yet available in the CCD.
b. Unless otherwise indicated, the test-statistic calculated was a *t*-test of difference between means.
c. Some data were missing for the locale variables, due to nonresponse on the CCD, for 22 of the schools eligible for participation. No data were missing for the sample of schools randomly assigned to the study.
d. Chi-square test from contingency table of frequency counts for all levels of the locale variable.
e. *z*-test for proportions.
f. Some data were missing for these school average student characteristic variables, due to nonresponse on the CCD, for 97 of the schools eligible for participation. No data were missing for the sample of schools randomly assigned to the study.
g. Welch-Satterthwaite adjustment was used for unequal variances when calculating the *t*-statistic and associated *p*-value (unequal variance determined by Levene's test, $F = 5.04$, $p = .025$).
h. Due to small cell sizes, the average percentage of Asian students and average percentage of American Indian students were added to create a combined percentage of students who were either Asian or American Indian.
Source: Common Core of Data 2008/09 (NCES n.d.b).

Comparison of 65 analytic sample schools with 70 randomized schools

Five schools were lost to either withdrawal or merger after randomization. To examine differences that might have been introduced to the sample by these changes, the 65 schools in the analytic sample were compared with the original 70 schools on the same 12 baseline characteristics. There were no statistically significant differences between the two groups on any characteristic (table 2.11).

Table 2.11. Comparison of 65 analytic sample schools to 70 randomized schools on baseline characteristics for the impact year (2009/10)

School characteristic	Randomized sample (*n* = 70)	Analytic sample (*n* = 65)	Difference (SE)	Test statistic	*p*-value
Locale (percent)[a]				$\chi^2 = 0.07$[b]	.966
Urban	24.29	26.15	–1.87	–0.25[c]	.804
Suburban or Small City	50.00	49.23	0.77	0.09[c]	.929
Rural	25.71	24.62	1.10	0.15[c]	.884
School average student characteristics					
Number of students[d]	515.86 (277.92)	521.77 (278.82)	–5.91 (47.95)	-0.12	.902
Background characteristics of enrolled students					
Female (percent)[d]	49.01 (3.39)	48.92 (3.36)	0.09 (0.58)	0.16	.875
Racial/ethnic composition (percent)[d]					
White	58.33 (36.76)	56.79 (36.87)	1.53 (6.34)	0.24	.809
Black	25.69 (30.32)	26.86 (30.79)	–1.16 (5.26)	–0.22	.825
Hispanic	12.69 (12.95)	13.21 (13.09)	–0.52 (2.24)	–0.23	.818
Asian or American Indian	3.17 (5.39)	3.01 (5.33)	0.16 (0.92)	0.17	.868
Eligible for free or reduced-price lunch (percent)[d]	34.78 (29.12)	36.03 (29.79)	–1.25 (5.07)	–0.25	.806

a. Standard deviations are not reported because the percentage for each category was derived from a dichotomous variable.
b. Chi-square goodness of fit test.
c. z test for the difference between two proportions.
d. These data are provided in the Common Core of Data as the percentage of the school that is a certain demographic (for example, the percentage of the school that is female); therefore, a chi-square is not computed. Instead, a *t*-test is appropriate to compare schools on these averaged demographics.
Note: Percentages may not sum to 100 due to rounding. Standard deviations are in parentheses under the means for intervention and control schools where appropriate. A *t*-test was used for comparisons, unless otherwise noted.
Source: Common Core of Data 2008/09 (NCES n.d.b).

School characteristics and equivalence between conditions in randomized sample

While randomization balances groups on measured and unmeasured characteristics over many repeated samples, in any one sample there could be statistically significant differences between groups. Intervention and control schools were compared on the same schoolwide characteristics as above—those hypothesized to be at least moderately correlated with students' mathematics achievement—to examine whether there were any statistically significant differences between the intervention and control groups following random assignment (table 2.12). These student characteristics are not limited to grade 6. The schools in the study had different configurations: some were K–6, some were 6–8, and some were K–8.

Table 2.12. Schoolwide characteristics for intervention and control schools in the randomized sample

Schoolwide characteristic	Intervention (n = 36)	Control (n = 34)	Difference (SE)	Test statistic	p-value
Percentage of schools in each locale[a]				$\chi^2 = 5.65$[b]	.059
Urban	36.11	11.76	24.35	2.47[c]	.016
Suburban or Small City	41.67	58.82	-17.16	-1.46[c]	.145
Rural	22.22	29.41	–7.19	–0.68	.499
School average student characteristics					
Number of students	536.22 (272.50)	494.29 (286.03)	41.93 (66.76)	0.63	.532
Background characteristics of enrolled students[d]					
Female (percent)	48.20 (2.77)	49.87 (3.80)	–1.68 (0.79)	–2.12	.038
Racial/ethnic composition (percent)					
White	48.12 (38.45)	69.14 (31.98)	–21.02 (8.43)	–2.49[c]	.015
Black	34.34 (33.00)	16.53 (24.49)	17.81 (6.92)	2.58[c]	.012
Hispanic	14.35 (12.67)	10.93 (13.20)	3.42 (3.09)	1.11	.273
Asian or American Indian	3.02 (5.62)	3.33 (5.21)	–0.31 (1.30)	–0.24	.812
Eligible for free or reduced-price lunch (percent)	40.65 (30.98)	28.56 (26.03)	12.10 (6.86)	1.76	.082

a. Standard deviations are not reported because the percentage for each category was derived from a dichotomous variable.
b. Chi-squared goodness of fit test.
c. z test for the difference between two proportions.
d. These data are provided in the Common Core of Data as the percentage of the school that is related certain demographic (for example, the percent of the school student population that is female); therefore, a chi-square is not computed. Instead, a t-test is appropriate to compare schools on these averaged demographics.
Note: Percentages may not sum to 100 due to rounding. Standard deviations are in parentheses under the means for intervention and control schools where appropriate. A t-test was used for the comparisons, unless otherwise noted.
Source: Common Core of Data 2008/09 (NCES n.d.b).

Intervention schools were more likely to be in an urban area (36 percent) than were control schools (12 percent; $p = .016$). Also, on average, a lower percentage of intervention students were female (48 percent; control students, 50 percent; $p = .038$) and White (48 percent; control students, 69 percent; $p = .015$). Finally, a higher percentage of intervention students were Black (34 percent; control students, 17 percent; $p = .012$).

School characteristics and equivalence between conditions in analytic sample

The analysis was repeated for the 65 schools in the analytic sample (table 2.13). It was found that intervention schools were more likely to be in an urban area (37 percent) than were control schools (13 percent; $p = .026$). Also, on average, a higher percentage of intervention students were Black (35 percent; control students, 17 percent; $p = .014$), and a lower percentage of intervention students were White (47 percent; control students, 69 percent; $p = .016$). However, the previously noted statistically significant lower percentage of females in the intervention group in the randomized sample of 70 schools (see table 2.11) was no longer statistically significantly different in this analytic sample of 65 schools (48 percent; control students, 50 percent; $p = .060$).

Table 2.13. Baseline schoolwide characteristics for intervention and control schools in the analytic sample

Schoolwide characteristic	Intervention (*n* = 35)	Control (*n* = 30)	Difference (SE)	Test statistic	*p*-value
Percentage of schools in each locale[a]				$\chi^2 = 4.78$[b]	.092
Urban	37.14	13.33	23.81	2.29[c]	.026
Suburban or Small City	42.86	56.67	-13.81	-1.12 [c]	.262
Rural	20.00	30.00	–10.00	–0.93 [c]	.359
School average student characteristics					
Number of students	542.37 (273.93)	497.73 (287.19)	44.64 (69.95)	0.64	.524
Background characteristics of enrolled students					
Females (percent)[d]	48.19 (2.81)	49.76 (3.78)	–1.57 (0.82)	–1.92	.060
Racial/ethnic composition (percent)[d]					
White	46.69 (38.03)	68.58 (32.22)	–21.89 (8.83)	–2.48	.016
Black	35.30 (32.97)	17.01 (25.13)	18.29 (7.22)	2.53[e]	.014
Hispanic	14.75 (12.62)	11.41 (13.60)	3.35 (3.26)	1.03	.308
Asian or American Indian	3.10 (5.70)	2.90 (5.00)	0.002 (1.34)	0.12	.907
Eligible for free or reduced-price lunch (percent)[c]	41.31 (31.18)	29.86 (27.30)	11.44 (7.25)	1.56	.123

a Standard deviations are not reported because the percentage for each category was derived from a dichotomous variable.
b. Chi-squared goodness of fit test.
c. z test for the difference between two proportions.
d. These data are provided in the Common Core of Data as the percentage of the school that is a certain demographic (for example, the percentage of the school that is female); therefore, a chi-square is not computed. Instead, a *t*-test is appropriate to compare schools on these averaged demographics.
e. Satterthwaite correction was used to calculate *t*-values for these items to account for observed inequalities in variances across samples.
Note: Percentages may not sum to 100 due to rounding. Standard deviations are in parentheses under the means for intervention and control schools where appropriate. A *t*-test was used for the comparisons, unless otherwise noted.
Source: Common Core of Data 2008–09 (NCES n.d.b).

Teacher and student characteristics and equivalence between conditions in analytic sample

Teacher characteristics hypothesized to be at least moderately correlated with students' mathematics achievement were also compared, to examine whether there were any statistically significant differences between the intervention and control groups in the analytic sample. The teacher characteristics were collected from the teacher background survey administered to teachers in all study schools at the start of the impact year. The characteristics of participating grade 6 mathematics teachers were compared by group using weighted averages of these teacher characteristics at the school level to account for differences in the number of teachers across schools (table 2.14). The comparison shows that a lower percentage of teachers in intervention schools were White (82 percent) than those in control schools (95 percent, $p = .047$). It also showed that a lower percentage of teachers in intervention schools had majored in mathematics in college (1 percent) than those in control schools (14 percent, $p = .027$).

Table 2.14. Baseline characteristics of the teachers in the analytic sample aggregated at the school level

Study participant characteristic	Intervention (*n* = 35 schools; 72[a] teachers) (SD)	Control (n = 30 schools; 58 teachers) (SD)	Difference (SE)	Test statistic	*p*-value[b]
Teacher background					
Age (mean)	38.76 (10.41)	40.02 (11.68)	–1.26 (1.94)	–0.65	.520
Female (percent)	72.22 (45.10)	72.41 (45.09)	–0.19 (0.39)	–0.02	.981
Teacher ethnicity (percent)[c]					
White	81.91 (38.73)	95.03 (22.34)	–13.12 (0.71)	–2.03[d]	.047
Black and other	18.09 (38.74)	4.97 (22.34)	13.12 (0.71)	2.03	.047
Teacher years of experience (mean)					
At current school	7.01 (6.77)	7.53 (7.58)	–0.52 (1.40)	–0.37	.711
In current district	8.69 (8.80)	9.74 (8.82)	–1.05 (1.60)	–0.65	.516
Total teaching experience	11.08 (9.12)	11.80 (9.18)	–0.72 (1.67)	–0.43	.669
Teacher highest degree attained (percent)					
Bachelor's degree	55.86 (50.18)	55.66 (50.17)	0.20 (0.44)	0.02	.985
Graduate degree (Masters or PhD)[e]	44.14 (50.18)	44.34 (50.17)	–0.16 (0.44)	–0.02	.985
Teacher mathematics training (percent)					
Majored in mathematics	1.39 (11.79)	13.79 (34.78)	–12.40 (1.08)	–2.26[d]	.027
Number of upper division college mathematics courses (above calculus) taken	1.14 (2.04)	1.94 (2.77)	–0.80 (0.48)	–1.67	.100
Hours of PD in mathematics taken in past three years	41.16 (112.59)	42.21 (78.47)	–1.06 (20.42)	–0.05	.959
Used CMP or CMP2 previously (percent)	2.78 (16.55)	5.17 (22.34)	–2.39 (0.93)	–0.69	.490

a. Three teachers did not provide background data at baseline and were excluded from this analysis, reducing the *n* to 72.
b. The *p*-values were adjusted for clustering of teachers within schools.
c. Due to disclosure risk, Black, American Indian, Asian and Hispanic teachers were combined into “Black and other”
d. Satterthwaite correction was used to calculate *t*-values for these items to account for observed inequalities in variances across samples.

e. Due to disclosure risk, the categories of Masters and Ph.D. were combined.
Note: Standard deviations are in parentheses under the means for intervention and control schools. A *t*-test adjusted for clustering using a two-level multi-level model was used for the comparisons. Data on years of teaching experience were adjusted for teachers completing the survey at different points in the study.
Source: Teacher background survey.

Finally, the pretest assessment scores for the intervention students were compared with those of the control students, aggregated at the school level and compared across schools using weighted averages by assignment condition in the analytic sample (table 2.15). Students in intervention schools scored, on average, 14.09 points lower on the TerraNova pretest than students in control schools, and this unadjusted mean difference was statistically significant ($p = .003$).

Table 2.15. Baseline characteristics of students in the analytic sample, aggregated at the school level

Student pretest scores (mean)	Intervention (*n* = 35 schools) (SD)	Control (*n* = 30 schools) (SD)	Difference (SE)	Test statistic	*p*-value[a]
	(*n* = 2,810 students)	**(*n* = 2,678 students)**			
TerraNova basic battery pretest[b]	656.56 (20.48)	670.46 (16.06)	-13.90 (4.59)	-3.03	.004
	(*n* = 2,597 students)	(*n* = 2,544 students)			
Perceived task value pretest[c]	38.70 (2.38)	37.95 (2.02)	0.75 (0.51)	1.45	.144

a. The *p*-values were adjusted for clustering of students within schools
b. Scaled score of 0 to 800 with the following proficiency cutscores: progressing (649), nearing proficiency (677), proficient (709), and advanced (743; CTB/McGraw-Hill 2003).
c. Score of 7–49 on 7-item pretest.
Note: Standard deviations are in parentheses under the means for intervention and control schools. A *t*-test adjusted for clustering using a two-level multi-level model was used for comparisons.
Source: Data collected from students on the TerraNova and PTV administered by the study team.

To conclude, statistically significant differences between the intervention and control groups in the analytic sample were identified at baseline. These include TerraNova pretest scores, teachers who were mathematics majors, teacher and student ethnicity, and the location of the school. All these differences were statistically controlled for at the school level in the models used to estimate the impact of CMP2 on TerraNova and PTV scores, as described in the data analysis methods section of this chapter.

Data collection instruments

This section describes the instruments used to collect data on the variables of interest. These include instruments for measuring the primary and secondary outcomes, instruments for measuring relevant covariates, instruments for measuring implementation, and administrative data sources.

Primary and secondary outcomes

The primary and secondary outcomes were measured at pretest (baseline) and posttest using the mathematics subtest of the TerraNova and the Eccles and Wigfield PTV scale.

Math subtest of the TerraNova

The TerraNova CAT™2 Basic Multiple Assessments Form (CTB/McGraw-Hill 2003)—a standardized test with national norms—was used to collect both pretest and posttest data. Students have 90 minutes to complete the test. The TerraNova covers number and number relations; computation and numerical estimation; operation concepts; measurement; geometry and spatial sense; data analysis, statistics, and probability; patterns, functions, algebra; problem solving and reasoning; and communication. The TerraNova has a reliability of 0.91, based on a nationally representative sample (McGraw-Hill 2002). Scaled scores range from 0 to 800, with the following proficiency cutscores: progressing (649), nearing proficiency (677), proficient (709), and advanced (743; CTB/McGraw-Hill 2003).[34]

Eccles and Wigfield PTV scale

Eccles and Wigfield (1995) identified three domains of adolescent self and task perception of mathematics, perceived task value (PTV), perceived ability/expectancy, and perceived task difficulty. It was hypothesized that CMP2 was unlikely to impact grade 6 students' perceived ability/expectancy and perceived task difficulty during a single school year. Therefore, the PTV scale was selected as the sole secondary outcome and was used to collect both pretest and posttest data. This outcome is measured through 7 items of a 19-item survey (Eccles and Wigfield 1995; see appendix D). Each item is measured on a 7-point Likert-type scale (from 1 = not at all good to 7 = very good), providing a maximum score of 49 points.

At the broadest level, the PTV construct measures the value and importance students place on mathematics. According to Eccles and Wigfield (1995), PTV comprises three distinct subconstructs:

- Intrinsic interest value: inherent enjoyment the student derives from engaging in mathematical tasks (two items, internal consistency alpha of 0.76).
- Attainment value/importance: importance of doing well in mathematics relative to the student's self-schema and personal values (three items, alpha of 0.70).
- Extrinsic utility value: the value the task acquires because it is instrumental in reaching a variety of long- and short-range goals (two items, alpha of 0.62).

Eccles and Wigfield (1995) uncovered the PTV factor through exploratory factor analysis and confirmed its structure through a confirmatory factor analysis. However, their factor analysis was conducted on a student sample that comprised White students in grades 5–12. The student sample for the current study is diverse in race/ethnicity and restricted to grade 6. Therefore, before including student scores for the PTV construct, a confirmatory factor analysis was

[34] CTB enlisted more than 50 experienced teachers and curriculum experts to determine the levels of proficiency and scaled score cutscores for these levels in 1996. These cutscores and proficiency levels have been used ever since and are provided in the TerraNova technical report (CTB/McGraw-Hill 2003).

conducted using the pretest sample. The factor structure was confirmed and the seven-item PTV scale was reliable with a Cronbach's alpha of 0.76, which is considered acceptable for research purposes (Henson 2001).

Relevant covariates

While many school-level covariates were available through the CCD, current teacher-level data could be obtained only by using a teacher background survey.

Teacher background survey

A 10-question survey was used to collect data on teachers' prior use of CMP and CMP2, prior teaching experiences, college degrees, mathematics PD, and demographic characteristics (appendix E). Any statistically significant differences between intervention and control schools on these characteristics in the analytic sample were controlled for when estimating the impact of CMP2.

When the study team was notified of a change in teachers, a package including a letter of introduction to the study, the teacher background survey, and a consent form was sent to the school for the new teacher(s) to complete. The data for teachers who completed the survey during the implementation year were adjusted by adding one year to any response related to years so that their responses could be compared with teachers completing the survey during the impact year.[35] Teachers' ages as of September 1, 2009, were calculated based on their reported birthdates. Any change in teachers was considered a part of a "typical" CMP2 implementation.

Measures of implementation

The study was designed to gather CMP 2 implementation data in intervention schools at the classroom level and observational data on activity in control schools. This was accomplished through use of PD participation records, an intervention teacher monthly online survey, an intervention teacher end-of-year survey, and classroom observation protocols.

Professional development participation

Participation in the recommended PD sessions was measured with trainer attendance records, which were cross-referenced in discussions with intervention teachers during classroom observations.

Intervention teacher monthly online survey

A monthly online survey was designed to measure reported progress in the curriculum (see appendix E). Intervention teachers were asked to complete one survey each month, even if they taught multiple sections of grade 6 mathematics using CMP2.36

[35] A year was added to teacher responses for years teaching in the school, years teaching in the district, and total years teaching. Number of PD hours in the last three years was not used in the baseline comparison because the data gathered for the implementation year could not be adjusted to compare with teachers who responded during the impact year.
[36] Forty-eight percent of the intervention teachers (36 of 75) taught more than one section of grade 6 mathematics. (Data on number of class sections were not available for the seven intervention teachers not implementing CMP2.)

The survey asked teachers three questions during the implementation year: which CMP2 units they had completed that month, what went well that month, and what they had difficulty with in the past month of implementing CMP2.

Two questions were added during the impact year. To address whether intervention teachers were meeting the standard of 50 minutes a day suggested by the publisher, teachers were asked to estimate the time they spent per week on CMP2 instruction by selecting from time ranges (for example, 2–3 hours). Teachers were asked about hours per week, rather than minutes per day, because it would be easier for teachers following a weekly schedule that did not involve mathematics instruction every day. These data were important because instructional time has been shown to be related to student performance (Suarez et al. 1991).

Finally, teachers indicated which (if any) supplemental materials they had used by selecting from a list of eight types of supplemental materials (for a list of these eight types, see question 5 of the monthly online survey in appendix E).

Intervention teacher end-of-year survey

Intervention teachers were also asked to complete an end-of-year survey (see appendix E). Teachers reported the units completed and estimated the average time per week that they spent on mathematics instruction. They were also given a list of supplemental materials and asked to indicate which kinds they had used.

Classroom observation protocols

Lab personnel observed each intervention classroom twice during the implementation year and all classrooms (intervention and control) twice during the impact year (fall 2009 and spring 2010). The observations were designed to gather objective data on the implementation of CMP2 in the intervention schools and on whether the control schools were using CMP2 and CMP2-like practices.

Because a written protocol was desirable to increase standardization across observers, the protocol developed for intervention school observations in Eddy et al. (2008) was adapted for use in control schools, with the protocol publisher's approval. As a result, two forms of the protocol were available to observe teacher practices, one for intervention classrooms and one for control classrooms. These protocols align with the theory of change model presented earlier in this report (see figure 1.1). Detailed information, including the parallel structure of the two protocols and observer training data, is in appendix F.

The classroom observation protocol documented the time of day the class period was observed, length of the class period, allocation of time to different instructional activities, teacher practices and strategies, student practices, curriculum materials in use, and any forms of assessment, feedback, or grading observed during the visit. There were also several indicators of CMP2-like activity to be recorded. The following is a description of the indicators and the number of points possible for each:

These teachers responded to the monthly online survey by summarizing their experiences with the curriculum across their class sections. This instruction was not printed on the survey but was clarified with the teachers by lab personnel.

- *Making connections (five points).* This teacher practice indicator is based on five items that document evidence of teachers:
  - Connecting concepts taught in class to things students already know.
  - Making connections to the real world.
  - Using alternative teaching strategies to help make connections.
  - Assessing students' prior knowledge to make connections to new concepts.
  - Referencing a real-world connection during the first part of the lesson or introduction to the mathematics activity.

- *Teacher factors related to student responsibility for learning and complex thinking (11 points).* This teacher practice indicator is based on seven items that document evidence that:
  - Students engage in complex thinking.
  - Classroom seating is conducive to group or pair work.
  - The teacher is more of a "guide on the side" than a "sage on the stage."(five points)
  - The teacher explains learning goals of the lesson.
  - The teacher expects students to answer each other's questions.
  - The teacher creates an environment where students are expected to work with and help each other.
  - The teacher encourages curiosity and creativity.

- *Student evidence of taking responsibility for learning and complex thinking in class discussion (five points).* This student practice measure is based on five items that document evidence that students are:
  - Answering each other's questions.
  - Making connections to previous lessons.
  - Introducing more than one way to solve a problem.
  - Taking turns to answer teacher probes.
  - Collaborating with other students to solve a problem.
- *Student evidence of taking responsibility for learning and complex thinking in groups/pairs (five points).* This practice is based on the same five items as above but is focused specifically on student behavior while working in groups/pairs.

- *Time on practices more and less like CMP2.* The classroom observers documented the number of minutes for the class period observed. The observers then documented the number of minutes spent on the following practices that are more like CMP2 practices: class discussion, small group work, pair work. They also documented the number of minutes spent on the following practices that are less like CMP2 practices: lecture and independent work.

These activities are represented as percentages of the class time observed. The remaining percentage was spent on activities common to any class section: taking attendance, assigning homework, and the like.

Administrative data sources

Class rosters

Class rosters included teacher and student enrollment data. The rosters were used to confirm student participation in general education grade 6 mathematics classes and to check potential teacher mobility between intervention and control schools.

Data collection methods

This section describes the procedures for collecting data using each instrument.

TerraNova and PTV

Student participation in testing

Students were considered eligible to participate in testing if they were enrolled in a regular grade 6 class section, agreed to participate via signed assent, and had not been removed from the study by parent refusal to consent. Teachers administered the tests under supervision by trained study team members. Procedural testing decisions—such as time and testing site—were left up to each school.

Accommodations

The general rule followed for testing was to provide students with any special accommodations requested by the school principal and confirmed by the classroom teacher. These were the same accommodations students were to be provided for state testing based on their individual education plans. English language learner students provided with Spanish language translation for their state test were able to take the TerraNova test and PTV survey in Spanish upon school request. No students took the TerraNova in Spanish for the pretest and no students took the PTV survey in Spanish for either the pretest or the posttest. Five students completed a TerraNova test in Spanish for the posttest.[37]

Testing environment

The TerraNova pre- and posttest were administered in similar settings, such as a quiet auditorium or cafeteria, in both intervention and control schools. At each school, study team members provided teachers with a brief training on test administration, following written guidelines prepared by the principal investigators. Teachers administered the student informed-assent forms and tests according to these guidelines. The full TerraNova assessment required 90 minutes to administer.

[37] CTB/McGraw-Hill provides a Spanish language version of the TerraNova, sold under the product name SUPERA®, which was made available for students needing a Spanish language assessment. The Spanish version was psychometrically validated using similar procedures as the English version. Scores derived from the Spanish version were analyzed along with those obtained through the English version.

PTV schedule

The PTV inventory required 15 minutes. Schools were expected to administer the PTV the same day as the TerraNova, but those unable to comply were given the opportunity to administer the PTV up to one week in advance of the TerraNova.

Make-up tests

To the extent that the schools allowed, study team members conducted make-up sessions in any schools where an eligible student did not participate in testing.

Teacher background survey

Intervention classroom teachers completed the teacher background survey during the summer 2008 PD sessions. The surveys were mailed to the control classroom teachers and any new intervention teachers not present during the implementation year. The surveys were collected in the schools during the pretesting sessions in September and October 2009.

Professional development participation

The publisher's trainers submitted attendance documentation to the study team. Lab personnel supplemented this information through conversations with intervention teachers during classroom observations and reported their findings to the study team.

Intervention teacher monthly online survey

An email with a link to the monthly online survey was sent to each participating teacher during the first week of each month. If a teacher failed to respond, additional emails were sent each week for two weeks following the initial mailing. In some cases, school firewalls or other software interfered with the receipt of the emails. Lab personnel worked with teachers to resolve delivery problems during the initial classroom observations.

Teacher name, school, e-mail, and responses were recorded in a database with a timestamp for the date started, the date completed, and the date modified.

Intervention teacher end-of-year survey

The cumulative end-of-year survey was administered to intervention teachers on the day of the posttest. For any teacher absent that day, the survey was left with the substitute and collected during make-up testing.[38]

Classroom observation protocols

Trained lab personnel scheduled all classroom observations directly with each school. Only one section was observed for each teacher, even if the teacher taught multiple sections. The observation was arranged to ensure, when possible, that multiple teachers at the same school could be observed on the same day. For consistency, the same mathematics class section observed in the fall was observed again in the spring.

[38] The seven intervention teachers who did not participate in the study did not complete the end-of-year survey.

Class rosters

Class rosters were obtained from each participating school before the beginning of each school year.

Data analysis methods

This section describes the data analysis methods used to evaluate CMP2 implementation and the analytic methods used to examine this study's research questions. Sensitivity analyses and the missing data approach selected for this study are also described.

CMP2 implementation

Data analysis plan for examining implementation

Teacher preparation was measured by PD participation and implementation was measured by teacher self-report and independent classroom observations.

According to the publisher, PD opportunities vary by the package purchased with the curriculum. This study used a typical package (five PD sessions). The publisher made no claims as to the minimum PD required for teacher effectiveness. Therefore, data on PD participation is used as part of the implementation analysis, but no statements can be made as to whether the observed participation rates were typical or that they had any impact on the findings.

The publisher did, however, recommend that a school spend 50 minutes per day (4 hours and 10 minutes per week) on mathematics instruction and complete at least six units of the CMP2 curriculum per year. Self-report data from the end-of-year teacher survey was used to determine the amount of time spent on mathematics per week for both the intervention and the control group, as well as the number of CMP2 units completed by intervention teachers.

Fidelity of implementation was measured using data from the classroom observations. Intervention and control classrooms were compared to explore whether differences in observed instructional practices existed at the time of observation. Data from observations of intervention classrooms were analyzed for evidence that CMP2-like activity was taking place. Similarly, data from the control classroom observations were reviewed for evidence that CMP2-like activity was not occurring.

Each type of data analysis is presented in the following section. Whenever two sets of implementation data were compared (for example, the same teachers over two time points or the intervention versus the control teachers), a test was run to gauge whether any difference was statistically significant, using $p < .05$ as the threshold for significance.

Participation in professional development

Five PD sessions were provided as part of CMP2 implementation.[39] The publisher's attendance data, verified and supplemented by teacher self-report, were examined. Descriptive data analyses

[39] As noted elsewhere, all new CMP2 teachers were offered two PD days in the impact year, but the additional three days were offered only if there were no experienced teachers at their school to act as mentor. This is the publisher's typical approach to PD.

were conducted and attendance numbers tabulated. The seven intervention teachers not using CMP2 were assigned zeros for the number of days of PD attended.

Total number of CMP2 units completed by intervention teachers

Although the monthly online surveys had response rates of 79–90 percent, the study team wanted to use more complete information and thus used the end-of-year survey (100 percent response rate). The responses to the end-of-year survey on the number of units covered during the school year were tabulated to determine the distribution of number of units completed, the average number covered, and the percentage of teachers who met the benchmark expectation of covering six units or more. The seven intervention teachers who did not use CMP2 were assigned zeros for the number of units completed for both the implementation and the impact years.

Time spent weekly on CMP2 by intervention teachers

The responses to the end-of-year survey on time spent weekly on CMP2 were tabulated into a frequency distribution to find out what percentage of teachers met the recommended level of 4 hours and 10 minutes per week. The seven intervention teachers who did not use CMP2 were assigned zeros for the time spent weekly on CMP2.

Classroom observations for intervention and control teachers

Classroom observation data were tabulated to generate a score for each observed teacher on the two teacher and two student practice measures. Observation data on how teachers spent their instructional time were compared with the total length of the session observed, to calculate the percentage of class time spent on activities that were CMP2-like practices and on activities that were not CMP2-like practices.

The number of points each intervention teacher received for implementing each area of practice observed was averaged across teachers by condition. To explore the degree to which CMP2 was associated with a contrast in instruction, intervention teachers' average measures during the impact year were compared with those of control teachers. This analysis was done separately for fall and spring observations. Mean observed practice scores and mean percentage of time measures were then averaged across intervention and control teachers who participated in classroom observations for each time point they were measured (fall 2009 and spring 2010), using an HLM to account for clustering of teachers within schools, with each area of practice as the outcome variable. See appendix G for additional information on the HLM and chapter 3 for the findings from this comparison.

Time spent weekly on mathematics by intervention and control teachers

The amount of time teachers reported spending on mathematics instruction each week was compared by assignment condition. Intervention and control teachers' self-reports of average total time spent on mathematics per week, made during spring observations, were averaged separately for intervention and control teachers and compared using an HLM to account for clustering of teachers within schools. Since these data were gathered during the spring classroom observation of the impact year, there is no data for the seven intervention teachers who did not participate in the classroom observations.

Diffusion of CMP2 into control schools

Diffusion was explored in three ways. First, class rosters were examined to check whether any intervention teachers crossed over into control schools. Second, teacher background survey responses were examined to determine how many control teachers, if any, had prior CMP2 experience. Finally, classroom observation data were reviewed to determine whether any control teachers were using CMP2. Control teachers' fall and spring responses to classroom observers' questions were tabulated to determine how many different curricula were reported as being in use in control schools during the impact year, as well as the number and percentage of teachers using each curriculum.

Effects of CMP2 on TerraNova and PTV outcomes

Students attending the same school are more likely to have similar mathematics achievement and PTV scores than students in different schools, since they are exposed to the same curricula, similar teachers, and other common school resources and policies. HLMs take the nested structure of the data into account by allowing correlated errors, thus generating more accurate standard errors and resulting in correct statistical inferences (Raudenbush and Bryk 2002).

The adjusted impact of CMP2 on the TerraNova and PTV was estimated using a two-level HLM, which accounted for the nesting of students (level-1 units) in schools (level-2 units). This benchmark model included an indicator of intervention/control group assignment, student pretest scores, a missing pretest score indicator, and covariates with statistically significant differences at $p < .05$) between the intervention and control schools on baseline characteristics.

In addition to the statistical significance of the effects of CMP2, the magnitude of the effects was expressed in standard deviation units. Specifically, the effect size was computed as a standardized mean difference (Hedges' g) by dividing the adjusted group-mean difference by the unadjusted pooled standard deviation of the student-level outcome measure, as recommended in the WWC Procedures and Standards Handbook (version 2.1), appendix B (p.45).

Sensitivity analyses

Benchmark models were altered to test how sensitive the impact estimates were to alternative model specifications and assumptions about missing data. Four sensitivity analyses were conducted:

- Unadjusted mean differences between intervention and control schools (that is, excluding covariates from the models).

- Handling missing data using case deletion instead of the dummy variable adjustment approach.

- Adjusted mean differences between intervention and control schools estimated using the benchmark HLM with two additional covariates (percent of students who were female and percent of teachers who were black) that exhibited pretest differences between intervention and control schools at $p < .10$.

- An alternative model specification for the TerraNova outcome measure using three levels instead of two, with teacher/classroom level included in addition to student- and school-levels.

Missing data for outcomes and covariates

Any observations with missing data on the TerraNova or PTV posttest were deleted from the analysis.

Missing data for covariates were addressed by applying the dummy variable adjustment technique, an effective way of dealing with missing data given the conditions that apply to this study (moderate rates of missing data overall and higher rates of missing data on the pretest than the posttest; Puma et al. 2009). The dummy variable adjustment retains all students with a missing pretest score but a nonmissing posttest score in the impact analysis.

Levels of missing data

The amount of missing data at pretest for the TerraNova was 202 students (120 intervention and 82 control), 4 percent of the analytic sample. The amount of missing data at pretest for the PTV was 541 students (335 intervention and 206 control), 10 percent of the analytic sample. At posttest, the amount of missing data was less than 1 percent for each outcome measure, and all students had data for at least one posttest outcome measure.

Data missing at random on pretest

Data were assumed to be missing at random on the pretest, meaning that the missing values on the pretest were unrelated to observed values on the pretest but could be related to the observed values of other covariates used to estimate the impact of CMP2 (Enders 2010).

Dummy variable adjustment

Under the missing-at-random assumption, Puma et al. (2009) have shown that, while the dummy variable adjustment for missing covariates may bias the covariates in regression analysis of observation data, it does not bias the independent variable (CMP2 in this case) in a cluster randomized trial. The manipulation of the independent variable—through random assignment—ensures on expectation (or many repeated trials) that the independent variable does not correlate with observed and unobserved covariates in the models.

Sensitivity analysis for missing data

Casewise deletion was used to test how sensitive the CMP2 impact estimate was to the use of the dummy variable adjustment approach to missing data. This method can be used for any type of statistical analysis, as no special computational methods are needed and, most importantly, bias is often minimal when pretest variables are included in the model as covariates (Graham 2009).

3. Examining Implementation

This chapter presents information on the implementation of CMP2 in intervention schools. This includes data on teacher PD participation, comparison of performance against the publisher's guidelines, classroom observations, and diffusion—to set the context for the impact results that follow in chapter 4.

Participation in professional development

The PD package included with this implementation of CMP2 consists of five PD sessions (days) for each new teacher in the implementation year. Two sessions were held the summer before the school year, and three were scheduled during the school year.

All teachers at intervention schools were offered the two summer sessions, regardless of whether they were present when the study began. Teachers who did not join their schools until after the implementation year were expected to receive mentoring from experienced colleagues in addition to the two summer sessions. If an experienced mentor was not available, these new teachers were offered the additional three days. Of the 75 teachers in the study for the impact year, 7 chose not to implement CMP2, 52 were trained in the implementation year (69 percent), and 16 were new to the study (21 percent). Ten of the new teachers were mentored in the use of CMP2 by previously trained sixth-grade teachers in their schools, which was the typical approach according to the publishers. The remaining 6 teachers received the 3 days of formal PD because there was no mentor teacher available in their school.[40]

Ideally, all teachers invited to PD sessions would attend all the sessions, but this did not happen. Just 53 percent of the teachers at intervention schools during the impact year attended five days of PD (table 3.1). All but the seven nonparticipating intervention teachers attended at least one PD session, and 71 percent attended at least three.

Table 3.1. Total number of days of CMP2 PD attended by intervention teachers

	Impact year teachers (*n* = 75)	
Number of PD days	**Percentage**	**Cumulative percentage**
0	9	9
1–2	20	29
3–4	18	46
5	53	100

Note: Includes the seven intervention teachers who did not implement CMP2 and completed zero units.
Source: For PD days one and two, records from the publisher were confirmed by teacher self-report. For the remaining days, data were collected solely from teacher self-reports.

[40] This RCT was framed as an effectiveness trial. As such, it was a test of the effect of CMP2 on student outcomes under implementation conditions that were *typical* for the sample of schools in the study, rather than conditions that may have been considered optimal. Teacher turnover occurred at the end of the first year of the study, leading to the need for new second year teachers to be trained by peer mentors rather than by the publisher's trainers. The publisher confirmed that this is what happens in typical implementations of CMP2.

The study team did not intervene to improve attendance. It is unknown how the PD attendance rates observed in this study compare with those in other places that have implemented CMP2.

Completion of CMP2 instructional units

Number of CMP2 units completed

Nearly two-thirds (64 percent) of the intervention teachers met the implementation benchmark of completing six or more units of CMP2 during the impact year (table 3.2). About one-fourth (27 percent) completed between one and five units.

Compliance with this publisher expectation was higher in the impact year than in the implementation year, when only 37 percent of intervention teachers completed six or more units.

Table 3.2. Number of CMP2 units completed, by study year (percent)

Number of units completed	Implementation year teachers(*n* = 82)	Impact year teachers		
		All teachers(*n* = 75)	Returning teachers (*n* = 59)	New teachers(*n* = 16)
0	9	9	12	0
1–5	55	27	24	38
6 or more	37	64	64	63

Note: This table includes the seven intervention teachers who did not implement CMP2 and completed zero units. Columns may not total 100 percent due to rounding.
Source: Teacher self-report data from monthly and end-of-year surveys.

Time spent weekly on CMP2

Sixty-eight percent of intervention teachers reported using CMP2 for four or more hours per week during the impact year (table 3.3), approximately equivalent to the publisher's recommended amount of time needed to complete the curriculum, (50 minutes per day over a five-day school week, equaling 4 hours and 10 minutes per week).

Table 3.3 Average hours spent on CMP2 each week during the impact year as reported by intervention teachers

Approximate amount of time spent on CMP2 each week	Percentage of teachers (*n* = 75)
0	9
Less than 2	10
2–4	12
4–5	17
5–6	24
6–7	11
7 or more	16
Total	99

Note: This analysis includes the seven intervention teachers who did not implement CMP2 and completed zero units. Total does not equal 100 percent due to rounding.
Source: End-of-year survey.

Instruction by intervention and control teachers

This section presents findings from analyses of classroom observation data. The first two analyses compare the instructional practices of intervention and control teachers during the impact year for the fall and spring semesters. These analyses are important because they could illuminate whether adopting CMP2 was associated with adopting various teaching practices; such differences could help explain any impacts on mathematics achievement or PTV. The two time points are examined separately, rather than together, to allow for the possibility that aspects of instructional practices vary over time. The final analysis compares intervention and control teachers on the average hours per week they reported spending on mathematics instruction during the impact year.[41]

Comparison of intervention and control teachers' instructional practices

Intervention and control teachers' instructional practices in the fall of the impact year differed significantly on four of the six measures (table 3.4). Intervention teachers received 2.89 more points (of 11) on the indicator of teacher factors related to student responsibility for learning and complex thinking (p = .000) and 1.81 more points (of 5) on the indicator of students taking responsibility for learning or complex thinking in groups or pairs (p = .000). They also spent more of their class time on activities considered CMP2-like (28 percent more; p = .000) and less of their class time on activities not considered CMP2-like (16 percent less; p = .002). See appendix H for detailed information on the coding of the observation protocol.

[41] For each of these analyses, we were unable to observe the classrooms of the seven intervention teachers who did not implement CMP2. The data for these teachers are considered missing. Therefore, caution should be taken when interpreting the findings of these analyses, as they pertain not to all intervention teachers but just to those who used CMP2.

Table 3.4. Comparison of CMP2 instructional practices observed in intervention and control teachers' classrooms, fall of the impact year

Instructional practices	Intervention teachers (*n* = 68) Mean points (SD)	Control teachers (*n* = 58) Mean Points (SD)	Difference (SE)	Percentage difference out of points possible	t-statistic	p-value
Teacher practices						
Making connections[a] (5 points possible)	2.41 (1.39)	2.04 (1.40)	0.37 (0.28)	7.40	1.31	.196
Teacher factors related to student responsibility for learning and complex thinking[b] (11 points possible)	7.89 (2.33)	5.00 (2.25)	2.89 (0.44)	26.27	6.51	.000
Student practices[c]						
Student evidence of responsibility for learning and complex thinking *in class discussion* (5 points possible)	3.08 (1.68)	2.52 (1.44)	0.56 (0.37)	11.20	1.52	.134
Student evidence of responsibility for learning and complex thinking *in groups/pairs* (5 points possible)	3.93 (1.64)	2.12 (2.07)	1.81 (0.35)	36.20	5.25	.000
Time (percent of total class period observed)[d]						
Percent of class time on activities *more* like CMP2 practices	72.49 (22.57)	44.43 (29.60)	28.06 (5.39)	—	5.21	.000
Percent of class time on activities *less* like CMP2 practices	11.68 (17.54)	27.52 (26.96)	–15.84 (4.78)	—	–3.31	.002

— is not applicable (no total possible points).

a. Measured by five items that document evidence of teachers connecting concepts taught in class to things students already know, making connections to the real world, using alternative teaching strategies to help make connections, assessing students' prior knowledge to make connections to the new concepts, and referencing a real-world connection during the first part of the lesson or introduction to the mathematics activity.
b. Includes seven items that gauge whether teachers expect students to engage in complex thinking, arrange classroom seating conducive to group or pair work, use more of a "guide on the side" versus "sage on the stage" pedagogy (five-point item), explain the learning goals of the lesson, expect students to answer each other's questions, create an environment where students are expected to work with and help each other, and encourage curiosity and creativity.
c. Includes answering each other's questions, making connections to previous lessons, introducing more than one way to solve a problem, taking turns to answer teacher probes, and collaborating with other students to solve a problem.
d. Classroom observers documented the total number of minutes for the class period observed, the number of minutes spent on practices considered more like CMP2 practices (class discussion, small group work, pair work), and the number of minutes spent on the practices considered less like CMP2 practices (lecture and independent work).
Note: This table excludes the seven intervention teachers who did not implement CMP2; they did not participate in classroom observations. The inferential statistics (standard error, *t*-statistic, and *p*-value) were adjusted for teachers nested within schools by estimating a two-level model, HLM 6.0, with the indicator variable for school group (intervention versus control) at level 2. Standard deviations for the intervention and control group means were not adjusted for clustering and were calculated using an independent samples *t*-test in SPSS.
Source: Classroom observation protocol. See appendix H for detailed information on the coding of the observation protocol.

The differences on these four measures were also statistically significant in the spring of the impact year (table 3.5). In addition, the two groups also differed by a statistically significant margin on the other two spring classroom observation measures. Intervention teachers received 0.58 more points (of 5) on the indicator for making connections ($p = .024$) and 1.00 more points (of 5) on the indicator of students taking responsibility for learning or complex thinking in class discussions ($p = .004$). Thus, intervention and control teachers' instructional practices in spring 2010 differed significantly on all six measures, suggesting a clear distinction in classroom practices associated with adopting CMP2. Additional information about the observation measure is included in appendix H.

Table 3.5. Comparison of instructional practices observed in intervention and control teachers' classrooms, spring of the impact year

Instructional practices	Intervention teachers (n = 68) Mean points (SD)	Control teachers (n = 58) Mean points (SD)	Difference (SE)	Percentage difference out of points possible	t-statistic	p-value
Teacher practices						
Making connections[a] (5 points possible)	2.65 (1.18)	2.07 (1.59)	0.58 (0.25)	11.60	2.32	.024
Teacher factors related to student responsibility for learning and complex thinking[b] (11 points possible)	8.23 (2.36)	4.99 (2.31)	3.24 (0.44)	29.45	7.33	.000
Student practices[c]						
Student evidence of responsibility for learning and complex thinking *in class discussion* (5 points possible)	3.15 (1.57)	2.15 (1.73)	1.00 (0.33)	20.00	3.02	.004
Student evidence of responsibility for learning and complex thinking *in groups/pairs* (5 points possible)	3.78 (1.79)	0.98 (1.78)	2.80 (0.33)	56.00	8.41	.000
Time (percent of total class period observed)[d]						
Percent of class time on activities *more* like CMP2 practices	70.60 (27.33)	36.88 (27.80)	33.72 (5.80)	—	5.81	.000
Percent of class time on activities *less* like CMP2 practices	12.76 (18.17)	26.47 (22.90)	-13.71 (4.05)	—	–3.38	.002

— is not applicable (no total possible points).

a. Measured by five items that document evidence of teachers connecting concepts taught in class to things students already know, making connections to the real world, using alternative teaching strategies to help make connections, assessing students' prior knowledge to make connections to the new concepts, and referencing a real-world connection during the first part of the lesson or introduction to the mathematics activity.

b. Includes seven items that gauge whether teachers expect students to engage in complex thinking, arrange classroom seating conducive to group or pair work, use more of a "guide on the side" versus "sage on the stage" pedagogy (five-point item), explain the learning goals of the lesson, expect students to answer each other's questions, create an environment where students are expected to work with and help each other, and encourage curiosity and creativity.

c. Includes answering each other's questions, making connections to previous lessons, introducing more than one way to solve a problem, taking turns to answer teacher probes, and collaborating with other students to solve a problem.
d. Classroom observers documented the total number of minutes for the class period observed, the number of minutes spent on practices considered *more* like CMP2 practices (class discussion, small group work, pair work), and the number of minutes spent on the practices considered *less* like CMP2 practices (lecture and independent work).
Note: This table excludes the seven intervention teachers who did not implement CMP2; they did not participate in classroom observations. The inferential statistics (standard error, t-statistic, and p-value) were adjusted for teachers nested within schools by estimating a two-level model, HLM 6.0, with the indicator variable for school group (intervention versus control) at level 2. Standard deviations for the intervention and control group means were not adjusted for clustering and were calculated using an independent samples t-test in SPSS.
Source: Classroom observation protocol. See appendix H for detailed information on the coding of the observation protocol.

Comparison of intervention and control teachers' time spent on mathematics instruction

Adoption of CMP2 was associated with spending significantly more time on mathematics instruction during the impact year. On average, intervention teachers (5.82 hours) reported spending 1.18 more hours (19 percent) per week (14 more minutes per day) on mathematics instruction than control teachers (4.64 hours; $p = .002$; table 3.6). Thus, any observed effect for CMP2 should be considered a combination of the curriculum and additional instructional time.

Table 3.6. Teachers' average hours per week of mathematics instruction in the impact year, by study condition

	Intervention teachers (*n* = 68) (SD)	Control teachers (*n* = 58) (SD)	Difference (SE)	*t*-statistic	*p*-value
Average hours per week	5.82 (1.64)	4.64 (1.47)	1.18 (0.37)	3.22	.002

Note: Analysis of time for mathematics reported weekly by teachers during two classroom observations, fall 2009 and spring 2010. The inferential statistics (standard error, *t*-statistic, and *p*-value) were adjusted for teachers nested within schools by estimating a two-level model using HLM 6.0, with the indicator variable for school group (intervention versus control) at level 2. Standard deviations for the intervention and control group means were not adjusted for clustering and were calculated using an independent samples *t*-test in SPSS. This table excludes the seven intervention teachers who did not participate in classroom observations.
Source: Teacher self-report recorded by lab personnel on the classroom observation protocol.

Diffusion of the intervention into control schools

Two of three analyses found no evidence of diffusion. First, an analysis of information from school rosters found no evidence of crossover from an intervention school to a control school. No teachers assigned to teach grade 6 mathematics in an intervention school moved to teach grade 6 mathematics in a control school.

Second, fall and spring observations found no evidence of CMP or CMP2 curriculum in use by teachers in control schools. Control teachers reported a variety of official curricula adopted in their schools, none of which were CMP2. Sixteen curricula were reportedly in use across the 30 control schools, none used by more than 13 percent of the teachers (table 3.7). No attempt was made to analyze the content of the control schools' curricula other than to confirm that CMP2 was not being used.

Table 3.7. Grade 6 mathematics curricula reported in use by control teachers

Sixth-grade mathematics publisher and curriculum	Percent of control teachers	Number of control teachers (*n* = 58)
Glencoe MacMillan McGraw Hill, Math Connects	6.9	4
Glencoe, Mathematics Applications and Concepts Course 1	6.9	4
Harcourt Brace, Math Advantage	< 7	< 4
Holt, Math Course 1	< 7	< 4
Houghton Mifflin, Math Grade 6	6.9	4
MacMillan McGraw Hill, Mathematics	12.1	7
McDougal Littell, Passport to Math	< 7	< 4
McDougall Littell, Math Course 1	< 7	< 4
McGraw-Hill Glencoe, Mathscape Course 1	6.9	4
Prentice Hall, Middle Grades Tools for Success	< 7	< 4
Prentice Hall, Middle School Math Course 1	8.6	5
Prentice Hall, Middle School Math Course 2	< 7	< 4
Saxon, Math Course 1	< 7	< 4
Scott Foresman Addison Wesley, Grade 6 Math	12.1	7
Scott Foresman Addison Wesley, EnVision Math	6.9	4
Wright Group/McGraw Hill, Everyday Math	8.6	5

Note: Specific numbers that include four or fewer teachers (and the related percentages) cannot be presented due to a potential disclosure risk. The curriculum reported in use in the fall of the impact year was confirmed to be in use in the spring of the impact year as the adopted text for the school- or teacher-selected book for the class section.
Source: Classroom observation protocol.

One analysis did raise concerns about possible diffusion of CMP2 into control classrooms. Some (fewer than four)[42] control teachers, each teaching in a different school during the impact year, reported having a year of CMP2 experience. It is conceivable that these teachers were using teaching practices learned from CMP2, even though they were not using it during the study. If so, based on their teaching assignments, this potential diffusion could have affected up to 8 percent of control classrooms (12 of 149) in the study and up to 7 percent of all control students (183 of 2,759) in the analytic sample.

[42] The exact number of control teachers with prior experience with CMP2 cannot be stated due to a potential disclosure risk.

Summary of the examination of implementation

More than half the intervention teachers followed publisher-recommended guidelines for PD (53 percent), curriculum coverage (64 percent), and time using CMP2 (68 percent).[43] Compliance could have been higher, in theory, but this study documented the reality of how implementation took place in typical conditions, without researcher interference. It is not known how the compliance levels observed in this study—on days of PD attended, CMP2 units covered during the school year, and hours per day using the curriculum—compare with levels achieved in other districts that have adopted the curriculum. Further, the importance of meeting publisher-recommended guidelines is unclear; this study does not investigate whether higher compliance is associated with better student outcomes.

Adoption of CMP2 was associated with differences in classroom practices. Compared with control teachers, intervention teachers were observed implementing more of the publisher's recommended instructional practices, such as encouraging students to take responsibility for learning and complex thinking and spending more time using instructional approaches that are consistent with CMP2 practices. If the classroom observations measured practices that align well with the goals of CMP2 and are associated with improved student mathematics achievement, then the observed differential conditions between intervention and control teachers would appear consistent with the possibility for CMP2 to have an effect. If differences had not been observed on these measures, then the chances of finding an effect would seem lower.

The 68 intervention teachers who used CMP2 during the impact year reported spending an average of about one hour and 10 minutes more per week (14 minutes per day) on mathematics instruction than the control teachers. Thus, any observed effect should be considered a result of both the CMP curriculum and the additional instructional time.

There was some evidence of possible diffusion of CMP2 into control classrooms. A few control teachers had CMP2 experience and might have used instructional approaches associated with it in their classrooms. These teachers taught 7 percent of the control students in the analytic sample. To an unknown degree, this situation could reduce the chances of detecting any effect that CMP2 might have on student achievement or PTV. On the other hand, no control teachers were observed using the CMP2 curriculum in their classrooms, and no intervention teachers crossed over into a control school to teach grade 6 mathematics.

[43] Some readers might be interested in an analysis of the costs of implementing CMP2, presented in appendix I.

4. Results: Impact of CMP2 on Student Outcomes

This chapter presents findings on the impact of CMP2 on TerraNova and PTV scores and the sensitivity analyses conducted to determine the extent to which the estimates were subject to assumptions.

Main analysis to estimate impact of CMP2 on student outcomes

Estimating impact

The impact of CMP2 on the TerraNova posttest scores was quantified as the difference between the covariate-adjusted group means of the intervention group and those of the control group. The estimated mean TerraNova posttest score for intervention schools was 682.76 points and that for control schools was 682.16 points, an estimated difference of 0.60 (table 4.1). This difference was not statistically significant ($p = .777$). Based on the magnitude of this adjusted difference and its lack of statistical significance, the conclusion is that CMP2 had no statistically discernable impact on TerraNova posttest scores.

Similar results were found for the impact of CMP2 on the PTV posttest scores, but caution should be used when interpreting these findings. Although the coefficient alpha for the PTV met standard levels of acceptance for research purposes (Cronbach's alpha of 0.76 at pretest), the pre- and posttest PTV scores were nearly uncorrelated for the control group ($r = 0.0545$; $p = 0.0066$), suggesting a lack of stability between the administrations. This lack of test-retest stability in the PTV scores suggests either that the instrument is not a reliable measure of PTV or that the PTV construct itself is not a stable trait.

The impact of CMP2 on the PTV posttest scores was also quantified as the difference between the two groups' covariate-adjusted group means. The estimated mean PTV posttest score for intervention schools was 37.32 points, and that for control schools was 36.67 points, an estimated difference of 0.65. This difference was not statistically significant ($p = .109$). Based on the magnitude of this adjusted difference and its lack of statistical significance, the conclusion is that CMP2 had no statistically discernable impact on PTV posttest scores.

Based on these covariate-adjusted impact estimates, the conclusion is that CMP2 did not have a statistically detectable effect on the TerraNova or PTV posttests. In other words, on the primary standardized mathematics outcome (the TerraNova), the grade 6 students in CMP2 schools did not perform statistically differently than grade 6 students in control schools using other curricula. See appendix J for the complete set of parameter estimates from the multilevel models (with the fixed and random effects).

Table 4.1. Estimated impact of CMP2 on student TerraNova posttest score and posttest PTV score

Outcome measure	Intervention group mean (*n* = 35)	Control group mean (*n* = 30)	Estimated difference (SE)	95 percent confidence interval	Effect size[a]	*p*-value
Primary outcome: TerraNova scores						
Adjusted for covariates	682.76	682.16	0.60 (2.12)	[–3.64, 4.84]	0.02	.777
Secondary outcome: PTV scores						
Adjusted for covariates	37.32	36.67	0.65 (0.40)	[–0.15, 1.45]	0.09	.109

a. Effect sizes were calculated using Hedge's *g*, consistent with the guidance in appendix B of the WWC Procedures and Standards Handbook (version 2.1). The mean difference is standardized by the unadjusted student-level pooled standard deviation of posttest scores. The unadjusted student-level standard deviations were 37.68 for the control group and 39.77 for the CMP2 group for the TerraNova posttest, and 7.41 for the control group and 7.23 for the CMP2 group for the PTV posttest.

Note: There were 2,923 intervention students and 2,754 control students in the analytic sample that took the TerraNova posttest, for a total of 5,677 students. There were 2,898 intervention students and 2,686 control students in the analytic sample that took the PTV posttest, for a total of 5,584 students. All the values in this table were estimated using a two-level HLM, which accounted for nesting of students within schools and controlled for student's pretest score, school mean pretest score, urban locale, percentage of White teachers in the school, percentage of teachers that majored in math, percentage of Black students in the school, and percentage of White students in the school.

Source: Student posttest scores on TerraNova and PTV measures.

Sensitivity analyses using alternative models

Five alternative models were estimated to test the sensitivity of the CMP2 impact estimate from the main model to the CMP2 impact estimate from the alternative models.

Unadjusted CMP2 impact

First, an alternative model without any of the covariates included in the main model was used to estimate the unadjusted difference in the mean TerraNova posttest score between intervention and control schools. The purpose was to show how sensitive the CMP2 impact estimate was to the covariate adjustments. This alternative model was then re-estimated with the PTV as the outcome variable.

The unadjusted estimated impact of CMP2 on the TerraNova posttest was –13.51 scale score points, favoring grade 6 students in control schools, and was statistically significant ($p = .010$). This result shows that the estimated impact of CMP2 (see table 4.1) was sensitive to covariate adjustments, including pretest scores. This sensitivity was expected because there was a statistically significant mean difference in the analytic sample on the TerraNova pretest of –14.09 scale score points ($p = .003$) favoring grade 6 students in control schools prior to delivery of CMP2. To compare the intervention and control schools on the mean TerraNova posttest scores without adjusting for this mean difference on TerraNova pretest scores would have biased the mean TerraNova posttest difference in favor of the control schools. By adjusting for this difference in pretest scores, and for statistically significant differences in schoolwide and school-

level characteristics, the impact of CMP2 on TerraNova posttest scores was estimated more accurately. [44]

In contrast, the unadjusted impact of CMP2 on PTV posttest was 0.59 scale score points, was not statistically significant (p = .102), and was similar in magnitude to the adjusted CMP2 impact estimated using the main model with the PTV as the outcome. See tables J1 and J3 in appendix J for the complete results.

CMP2 impact adjusted for urban locale

Whether the unadjusted estimated impact of CMP2 on the TerraNova of –13.51 scale score points, favoring grade 6 students in control schools, was due to the baseline imbalance in the proportion of urban schools in the intervention group was questioned. Accordingly, the alternative model was expanded without the covariates described above to include a dummy variable for urban locale. The purpose was to analyze how sensitive the unadjusted CMP2 impact estimate from the first sensitivity analysis was to a statistical control for urban locale. This sensitivity analysis—unlike the others in this section—was developed post-hoc and not prespecified in the study plan.

When the urban locale dummy variable was added, the estimated impact declined from –13.51 scale score points to –6.87—and was no longer statistically significant (p = .117). This result shows that the estimated impact of CMP2 (see table 4.1) was sensitive to the higher proportion of urban schools in the intervention group (see table 2.13 in chapter 2), to the point that the unadjusted mean difference on the TerraNova posttest scores that was once statistically significant (p = .010) was no longer so when urban locale was controlled for.

This alternative model was not re-estimated for the PTV because there were no pretest differences between intervention and control schools on the PTV.

Missing pretest data approach

An alternative model using case deletion was estimated, rather than the dummy variable adjustment technique used in the main model. The purpose was to test how sensitive the CMP2 impact estimate was to the use of case deletion as an alternative, and more conservative missing data technique.[45] This alternative model was then re-estimated with the PTV as an outcome variable.

Applying case deletion requires the exclusion of students without valid scores for both pretest and posttest. Of the 5,689 students eligible to complete the TerraNova pretest and posttest, 5,475 students (96 percent) completed both. As a result, 214 students had a missing value on either the pretest or posttest, or both, and were excluded when estimating the alternative model for this sensitivity analysis. Of the 5,689 total students eligible to complete the PTV pretest and posttest, 5,043 students (89 percent) completed both. As a result, 646 students had a missing value on either the pretest or posttest, or both, and were excluded when estimating the alternative model for this sensitivity analysis.

[44] See tables 2.11 and 2.12 in chapter 2 for a list of the covariates and their characteristics.
[45] Conservative in the sense that case deletion does not adjust for missing data through imputation and statistical adjustments as the dummy variable adjustment does.

For the TerraNova posttest scores, the CMP2 impact estimate of 0.88 scale score points and its *p*-value ($p = .662$) were similar to the corresponding estimates obtained from the main model that used the dummy variable adjustment to address missing data on the pretest. For the PTV posttest scores, the CMP2 impact estimate of 0.61 scale score points and its *p*-value ($p = .141$) were also similar to the corresponding estimates. For both outcomes, the impact of CMP2 remained less than one scale score point, and these impacts were not statistically significant.

Based on these results, the conclusion is that the CMP2 impact estimates (and standard errors) for the TerraNova and PTV were invariant to using case deletion. See appendix J for the complete set of results.

Controlling for group differences on covariates at $p < .10$

The adjusted CMP2 impact was estimated using an alternative model that included all the covariates in the main model plus two covariates that exhibited differences at pretest between intervention and control schools at $p < .10$. The purpose was to evaluate how sensitive the adjusted CMP2 impact estimate was to the inclusion of these covariates in addition to those already in the main model.

This sensitivity analysis was conducted by adding to the main model two variables that exhibited statistically significant between-group pretest differences ($p < .10$) at baseline in the analytic sample (see table 2.13 in chapter 2). The group difference in the variable of percentage of students who were female favored the control group (2 percent more; $p = .060$), and the variable of percentage of teachers who were Black favored the intervention group (11 percent more; $p = .063$). The alternative model was then estimated for both the TerraNova and the PTV.

For the TerraNova posttest, the estimated impact of CMP2 was less than one scale score point at –0.17 and was not statistically significant ($p = .938$). Similarly, for the PTV posttest, the impact of CMP2 was 0.53 scale score point and not statistically significant ($p = .221$). The conclusion is that the impact estimates for both student outcomes were insensitive to excluding the variables of percentage of teachers who were Black and percentage of students who were female when estimating the impact of CMP2 using the main model. See tables J1 and J3 in appendix J for the complete set of results.

Estimating three-level instead of two-level models

Finally, an alternative model was estimated using three levels (with classroom at level 2) instead of two levels to evaluate whether modeling the variance at level 2 affected the CMP2 impact estimate and its variance (or precision). Because this sensitivity analysis was purely methodological, and not substantive, this estimation process was not repeated for the PTV outcome.

To evaluate how sensitive the CMP2 impact estimate was to estimating a three-level model rather than a two-level one, the main model was re-estimated—but with a class section level with random effects inserted between the student and schools levels. The class section level represents the level in the nested structure of the data where there are multiple sections of students taught by the same teacher in the same subject. Thus, students are nested in class sections, which are nested in schools. The estimated impact of CMP2 was 1.05 scale score points and was not statistically significant ($p = .609$).

The conclusion is that the CMP2 impact estimate and its statistical significance for the TerraNova posttest scores were insensitive to the choice of estimating the main model using three levels rather than two. See table J2 in appendix J for the complete set of results.

Summary and conclusion

The magnitude and statistical significance of the CMP2 impact estimate was consistent across almost all the models. The scale-score difference on the TerraNova outcome between intervention and control schools was less than one point in most models, and these differences were not statistically significant at $p < .05$. The exception was the model estimated without covariate adjustments, where a difference of –13.51 was observed. This difference was statistically significant ($p = .010$).

This sensitivity of the CMP2 impact estimate to covariate adjustments was anticipated because there was a statistically significant difference between intervention and control schools on baseline covariates in the analytic sample. The main model adjusted for these differences; the alternative model did not.

In addition, the CMP2 impact estimates were consistent across all sensitivity analyses for the PTV. As with the primary outcome, there was no statistically discernable impact of CMP2 on the secondary outcome, the PTV.

5. Summary of Findings, Conclusions, and Study Limitations

The results of the examination indicated differences in the type of instructional activity taking place in intervention and control classrooms during observations and that the activity observed in the intervention classrooms was of the type expected in CMP2 implementation. Sixty-four percent of the intervention teachers reported implementing the curriculum at a level consistent with the publishers' recommended number of units completed, and 68 percent of the intervention teachers reported implementing the curriculum consistent with the recommended amount of class time per week.

Intervention teachers were observed spending statistically significantly more time on CMP2-like practices than control teachers (34 percent; $p < .001$). Students in intervention schools were more likely to be observed taking responsibility for their own learning and engaging in complex thinking ($p < .001$). These analyses did not include the seven intervention teachers who did not implement CMP2 and did not participate in classroom observations. Therefore, the results of the comparison of instructional practices should be interpreted with caution.

On average, intervention teachers reported spending 1.18 more hours of instructional time per week on mathematics than did control teachers ($p < .001$). This statistically significant difference should be considered when interpreting the findings.

The results also showed, however, that CMP2 as implemented in this study did not have a statistically significant effect on grade 6 mathematics achievement as measured by the TerraNova, which answered the primary research question.[46] The end-of-year difference between intervention and control schools on the TerraNova was 0.60 ($p = .777$). This translates to an effect size of 0.02, which is too small to be educationally meaningful. In sum, regular grade 6 mathematics students in intervention schools performed no better or worse on a standardized mathematics test than their peers in control schools. A majority (69 percent) of the intervention teachers were teaching CMP2 for their second year, so inexperience with the curriculum should not be a factor.

The results for the secondary research question were similar. The difference between intervention and control schools on the perceived task value of mathematics as measured by the PTV was 0.65 points ($p = .109$), or an effect size of 0.09. Thus, there was no statistically significant difference between groups on the PTV, and the small effect size is unlikely to be meaningful.

These results were insensitive to an alternative model specification with additional covariates (a three-level versus two-level HLM) and to an alternative approach to handling missing data through case deletion. The TerraNova results were sensitive to an unadjusted alternative model (without covariates), but the effect was likely due to a lack of equivalence at baseline. This lack of equivalence was also evident at random assignment, when comparisons of the original schools (n=70) on CCD school characteristics revealed a statistically significant difference in the proportion of urban schools, favoring the intervention group. There were no other differences at

[46] The primary research question was designed to test the impact of CMP2 on mathematics achievement. The secondary research question was exploratory. Thus, no adjustment for multiple comparisons was performed.

$p < .05$ on the CCD characteristics, including eligibility for free and reduced price meals. The imbalance was likely due to chance.

Attrition rates differed in intervention and control schools. The overall attrition rate across all schools was 7 percent. The differential attrition rate between intervention and control schools was 9 percent. Student-level attrition across all schools was 18 percent for the TerraNova and 20 percent for the PTV. Differential student attrition was 16 percent for TerraNova and 17 percent for the PTV. Student attrition was impacted by a loss of 761 students when schools were lost through withdrawal or merger.

The combination of overall attrition and differential attrition rates in the analytic sample is within the WWC boundaries at the school level, which is the level of random assignment. However, at the student level these rates are outside WWC boundaries, which suggests, according to the WWC simulations, that student level attrition could have biased the CMP2 impact estimate beyond the tolerable magnitude of 0.05 standard deviations.

School-level attrition is more concerning than student-level attrition because schools were the unit of random assignment, and the impact analysis measures CMP2 impact at the school level. A statistically significant difference between the attrition rates of intervention and control schools would cause concern about potential bias in the impact estimate.[47] This difference was not statistically significantly different ($p = .145$). However, the differential student attrition was statistically significant ($p < .001$) and should be considered.

The sensitivity analysis reported in Chapter 4, in which we used the TerraNova posttest score as the dependent variable and the TerraNova pretest score and urban locale as independent variables, showed that the observed pretest difference was associated primarily with the urban locale variable, because the pretest difference shrunk by half and was no longer statistically significant. From this we concluded that the urban locale variable, not attrition, was the primary source of the observed TerraNova pretest difference in the analytic sample.

The results of this study's impact analysis are consistent with Eddy et al. (2008), which found a positive but not statistically significant effect of CMP2 on grade 6 student achievement as measured by a standardized test. However, the two studies have important methodological differences. Eddy et al. (2008) randomly assigned teachers, rather than schools, to intervention and control conditions. Further, the sample size and hence statistical power for Eddy et al. (2008) was lower than the sample size and statistical power for this study. Finally, because teachers in Eddy et al. (2008) were randomly assigned within schools, there was a potential for diffusion of CMP2 instructional practices and curriculum use to control teachers and students that was not present in the current study.

This study also confirms similar null findings from the one large-scale quasi-experimental study that met WWC standards "with reservations" (Schneider 2000). That study also found that

---

[47] There is no available information in the literature on exactly how the combination of overall and differential student level attrition could bias the impact estimates, which makes it difficult to determine how large this bias, if it exists, could be. Because the sample is assumed to be sufficient to have a .80 chance of declaring an effect size of at least .20 standard deviations as statistically significant, the bias in the effect size would need to be approximately of that magnitude (given the observed effect size of .02 standard deviations) for this to occur.

students in CMP schools did not perform statistically significantly differently from students in control schools on a standardized measure of mathematics achievement. The improvement index was zero percentile points.[48]

Study limitations

The findings from this study should be interpreted in light of study limitations and considerations for generalization. Limitations include:

- The sample was unbalanced in the proportion of urban schools. A post-hoc analysis was conducted to test whether the unadjusted posttest difference between intervention and control schools on the TerraNova was associated with the higher proportion of urban schools in the intervention group. The unadjusted posttest difference of –13.51 scale score points was statistically significant ($p = .010$). When adjusted for the higher proportion of urban schools in the intervention group by including the urban locale variable in the analysis, the difference declined 49 percent to –6.87 scale score points and was not statistically significant ($p = .117$). Thus, the observed posttest difference on the TerraNova between the intervention and control groups was partially associated with the group imbalance on the urban locale variable.

- Regardless of source, this unbalanced sample may have resulted in biased impact estimates when using an unconditional model. Although the study team controlled for observed covariate differences in the benchmark model, it cannot be known whether all the bias that was potentially introduced by chance was eliminated. Future studies on CMP2 should include school locale (and prior achievement) as a blocking factor for random assignment

- The use of an implementation year, in theory, could threaten random assignment, since parents could choose schools based on knowing which were using CMP2 and which were not. However the extent to which this occurred and whether it occurred at all is unknown.

- Seven intervention teachers did not implement CMP2 and did not participate in teacher-level data collection activities beyond providing their background data. These teachers were included in the analyses of benchmarks for implementation (number of units completed, time spent weekly on CMP2) as intent-to-treat but not in the analyses of instructional practices or time spent on mathematics. Therefore, the findings based on these analyses should be considered with caution.

- Because this is an effectiveness study, we provide the percentage of teachers who met the publisher’s recommended number of units and amount of instructional time solely for information. As neither the developer nor the publisher was able to provide a definition of high or low fidelity of implementation, we relied on the developer’s implementation guide's benchmark for the recommended number of units to complete per year (6 units) and the recommended amount of time per day of instruction (50 minutes). No claim is made regarding a relationship between level of implementation and achievement.

[48] According to the WWC Procedures and Standards Handbook (What Works Clearinghouse 2008), the improvement index measures the difference between the percentile rank of the average student in the intervention condition and the percentile rank of the average student in the control condition.

- Three limitations relate to how data were collected through self-report surveys from the intervention teachers to analyze fidelity of implementation. First, the data could be inaccurate and slightly inflated, as intervention teachers may be attempting to appear compliant with the expectations and instructions provided by the trainers. Second, teachers teaching multiple class sections of grade 6 mathematics had to generalize across sections to report their responses, so their responses might not represent the actual level of implementation that each class received. And third, summarizing their responses from an entire year of activity could be inaccurate.

- Most of the observation protocol items used to measure implementation of CMP2 instructional practices and to compare the level of these practices between intervention and control teachers were yes/no variables that did not distinguish the extent (frequency/duration) to which each practice was used. Credit given to class sections where a practice (such as students introducing new ways to solve problems while in small groups) was observed just once, or briefly, equaled what was given to class sections where the same practice was observed repeatedly or for a long period of time. Credit does not characterize the extent to which a practice was observed, just whether it was observed or not. If there were systematic differences in the extent to which intervention or control teachers implemented certain practices, the observation protocol would not have been sensitive to them. Outcome measures might look similar for any two class sections or for any particular groups of class sections (including intervention versus control), but what occurred in the class section could have been different.

- Classroom observations were conducted twice each year. Findings based on the observational data are valid only for the time of observation, and should not be generalized over a longer period of time. In other words, just because a teacher was observed using CMP-like activities during the observation period does not insure that the same teacher performed that way for the entire school year.

- Observation sessions were scheduled ahead of time with each school, and observers were not blind to the study condition being observed, which is a potential threat to the internal validity of the measures.

- Although the coefficient alpha for the PTV met standard levels of acceptance (Cronbach's alpha of 0.76), the pre- and posttest PTV scores for control students were uncorrelated, suggesting a lack of stability between the pretest and posttest measures. The lack of test-retest reliability in the PTV scores suggests either that the instrument is not a valid measure of PTV or that the PTV construct is not a stable trait.

Considerations for generalization

Considerations for generalization include:

- This study used a volunteer sample. This sample is not necessarily a representative sample, and therefore the results may not generalize to other schools.

- This study included an implementation year and an impact year. Therefore, for the intervention group, 69 percent of the intervention teachers whose students took the TerraNova and PTV had been teaching CMP2 for two years. The results could have been different had effects been measured the first year of implementation.

- This study was conducted with the current version of CMP2. The results do not apply to other versions of CMP.

- This study compared mathematics achievement scores and PTV scores for students in intervention schools to that of students in schools with a variety of other curricula. Therefore, no conclusions can be drawn about how CMP2 might compare with any particular curriculum the control schools reported adopting.

- The conclusions drawn in this study about the effects of CMP2 on student math achievement are limited to student math achievement as measured by the TerraNova, and do not generalize to any other standardized test.

Appendix A. CMP2 Curriculum and PD

Sample CMP2 lesson

This section describes what a student could experience during a CMP2 lesson. Provided are examples of the content, student interactions, assessment, and teacher feedback associated with a lesson on common multiples and common factors. The example also includes questions teachers might ask during each phase of the instructional model. An overview of the goals and learning objectives of a lesson entitled "Riding Ferris Wheels" is in table A1.

Table A1. Selected lesson from CMP2

Unit: Prime Time	**Investigation 3: Common Multiples and Common Factors**
Lesson: Problem 3.1	Riding Ferris Wheels
Mathematical and Problem-Solving Goals	Recognize situations in which finding common multiples or common factors of whole numbers are important.
	Develop strategies for finding common multiples, common factors, least common multiple, and least common factor.
	Use patterns to reason and predict future occurrences and solve problems.

Source: Lappan et al. 2006b.

In this lesson, students investigate a situation in which a boy named Jeremy and his little sister Deborah are at a carnival. Each rides a different size Ferris wheel. Students are given the number of revolutions that each Ferris wheel makes in a certain amount of time. One Ferris wheel rotates every 20 seconds, and the other every 60 seconds. The students are also told that Jeremy and Deborah take off simultaneously, from the same initial starting position. Students then investigate how long it will take until both children reach the initial point at the bottom of the ride at the same time. Students explore finding common multiples of 20 and 60 to solve the problem. The following is a description of suggestions of what might regularly occur during the Launch, Explore, and Summarize phases of the instructional model. This information is provided in the teacher's guide for the unit (Lappan et al. 2006b, pp. 54–58).

Launch. The teacher asks the class questions to connect students' prior knowledge with the current topic of multiples and factors. The teacher also checks students' prior experience with the context of a carnival and a Ferris wheel ride, asking students who have ridden such a ride to share with the class how the ride works. To further launch the lesson, the teacher discusses with the students the impact of the size of the Ferris wheel on the experience of the rider.

Explore. For a short amount of time, students think individually about the Ferris wheel problem. They then explore the problem in small groups. During this time, the teacher circulates around the room and prompts groups having difficulty with questions such as, "After 20 seconds, where is the large Ferris wheel? The small Ferris wheel? Where are they after 40 seconds?" Students illustrate their strategies and solutions on poster paper.

Summarize. The student groups share their solution strategies with the class and justify how they reached their conclusions. The teacher questions students to ensure that the mathematical goals of the lesson are addressed (for example, "When is the least common multiple one of the numbers? When is it the product of the two numbers? When is it neither?").

After completing this problem, the students continue to find patterns and make predictions in the following scenarios: Cicada Cycles (Problem 3.2), Bagging Snacks (Problem 3.3), and Planning a Picnic (Problem 3.4). They then move to the next investigation: factorization of numbers.

Teachers are to implement each unit according to the guidelines in the teacher's guide. However, trainers work with schools to design plans to meet the needs of individual schools. For example, some schools have 90-minute class periods; others have 60-minute class periods. Some schools need to emphasize certain mathematical content areas more than others based on the needs of their students and their state mathematics standards. The trainers provide information on which lessons are optional so that the lessons of highest priority are completed prior to state testing.

Professional development

As part of the typical PD package provided by the publisher, teachers were offered five days of PD: two in the summer and three during the school year (table A-2). Prior to day three, a trainer visited each teacher's class section to tailor the PD for day three to meet the specific implementation needs of each intervention school.

Table A2. PD for CMP2

PD day	Description of PD content
Day 1 Summer 2008/09	Introduction to CMP2 components, implementation guide, pedagogy, and Prime Time unit
Day 2 Summer 2008/09	New Math discussion, introduction to Bits and Pieces 1 unit, Explore activities in Bits and Pieces I, instructional model (Launch, Explore, Summarize), and "I see, I think, I wonder" chart
Day 3 Fall 2008/09	What worked well and what didn't, continue exploration of Bits and Pieces I, Explore Shapes and Designs investigations, Guide to Class Observation, and using investigation as diagnostic tool
Day 4 Winter 2009/10	What worked well & what didn't, discussion around student work samples, Bits and Pieces II unit Exploration & Discussion to debrief what teachers have completed in the curriculum and to prepare for what is coming next
Day 5 Spring 2009/10	What worked well and what didn't, discuss area models, Covering and Surrounding unit Exploration and Discussion

Source: The publisher's training agenda provided during the PD (August 2008).

At the end of the implementation year, some intervention schools experienced changes in teachers due to turnover, reassignment of teachers to other positions, or rescheduling teacher assignments. Following the recommendation of the PD provider, the new teachers would not receive the three PD sessions during the year, which is considered typical for schools implementing CMP2. Instead, the teachers who had used the curriculum the previous year provided mentoring for any new teachers added to their schools. This is the typical approach to PD for schools purchasing the standard PD package. However, some new teachers did not have an experienced teacher at their school to mentor them in using CMP2. The publisher provided these new teachers with the opportunity to attend three PD sessions during the year in lieu of the mentoring, since it was not available at their school. PD was not provided during the impact year to teachers who had received it during the implementation year. A schedule of the PD activities for the teachers using CMP2 is shown in table A3.

Table A3. Schedule of PD and support activities for intervention teachers, implementation year and impact year

Participants	PD activity	Date window
Grade 6 mathematics teachers implementing CMP2, teaching assistants, special education and Title 1 teachers[a]	Initial training days 1 and 2	July 1–August 22, 2008
	Follow-up training day 3	November 3–28, 2008
	Follow-up training day 4	February 2–27, 2009
	Follow-up training day 5	April 6–30, 2009
New grade 6 mathematics teachers implementing CMP2, new teaching assistants, and new special education and Title 1 teachers	Initial training days 1 and 2	July 22–September 15, 2009
	Follow-up training teachers day 3[b]	November 10, 2009
	Follow-up training teachers day 4[b]	January 6, 2010
	Follow-up training teachers day 5[b]	February 24, 2010

a. Per the request of the participating schools, support teachers working with children included in the intervention class sections attended the PD to learn about the curriculum and instruction that would be implemented. This was explained as typical for any district PD provided to teachers to improve the collaboration between the classroom teacher and the support teachers.
b. Not offered to new teachers in schools that had experienced CMP2 teachers in 2009/10.
Source: Study records.

The trainers submitted attendance documentation for the first two days of PD. However, not all collected attendance data for the three days during the school year. Therefore, lab personnel verified the PD attendance submitted by the publisher and collected teacher self-report data on PD attendance for days 3–5 during the spring 2010 classroom observations using the form in table A4, customized for each school.

Table A4. Verification of PD attendance for intervention teachers

Class observer	School name	Lname	Fname	Summer 2008 day 1	Summer 2008 day 2	Fall 2008 day 3	Fall 2008 day 4	Spring 2009 day 5	Summer 2009 day 1 make-up	Summer 2009 day 2 make-up	Fall 2009 day 3 make-up	Fall 2009 day 4 make-up	Spring 2010 day 5 Make-up
Name	School X	Teacher A		Yes	No								
Name	School X	Teacher B		Yes	Yes								

Directions: Cross out any incorrect data and record the correct data. For example, if it says "No" but the teacher says they did attend PD, cross out "No" and write in "Yes."

Note: This document was customized for each school with the names of teachers for that school and the data provided by the publisher on PD attendance to verify PD attendance for all five days during the classroom observation.

Appendix B. Statistical Power Analysis as Conducted During the Design Phase

Assumed minimum detectable effect sizes for students' TerraNova and PTV scores

During the design phase of the study, a statistical power analysis was conducted to determine the number of schools, class sections, and students needed to detect a minimum detectable effect size (MDES) for the primary outcome of student mathematics achievement. The MDES is the smallest difference between the intervention and control groups on average student outcomes (measured in standard deviation units) that the study design could detect as statistically significant. This appendix describes the statistical power analysis laid out in the proposal for the design of this RCT.

As stated earlier, the school is the unit of assignment in this study. The school level is defined as all schools placed in the random assignment pool. Therefore, the study used a two-level model in our statistical power calculations that took into account clustering of students within schools (but ignored classroom-level clustering) when calculating statistical power. This approach is consistent with Schochet (2005):

> For school-based experimental evaluations, one design option is not to sample class sections within the intervention and control schools. For this option, either all relevant class sections in the selected schools are included in the research sample or students are assigned directly to the research sample without regard to the class sections they are in. (p. 21)

The assumptions regarding the magnitude of intraclass correlation coefficients (ICCs), or the proportion of variance in the outcome between schools as compared with the total variance in the outcome, were based on Schochet (2005). The study team assumed a level-2 ICC of 0.15 for the schools in the study, a value that Schochet used in presenting power estimates. When this study was designed in September 2006, there was little published information on unadjusted ICCs for studies in middle school mathematics. However, the longitudinal datasets examined and reported in table 2 of Schochet (2005, p. 23) had unadjusted ICCs that ranged from .10 to .20. With all other assumptions held constant, higher ICCs require more schools in the sample. Because of the cost of recruiting more schools at the highest unadjusted ICC, the study team assumed an unadjusted midpoint ICC of .15 for this study.

The TerraNova pretest of standardized mathematics achievement, administered by teachers and monitored by lab personnel at the beginning of the impact year, aggregated at the school level, was the school-level covariate in the power analysis. It was assumed that the covariate has a strong linear association with the outcome and that this association is similar within the intervention condition. Based on Bloom, Richburg-Hayes, and Black (2007), it was also assumed that the school-level pretest is as effective a covariate for school-level outcomes as a student-level pretest would be for student outcomes. Bloom, Richburg-Hayes, and Black (2007) show that the pretest R^2 can be 0.56 or higher at both the student and school levels. However, the study team conservatively assumed R^2 to be 0.50 for power calculations.

The study team used a school randomization design with students clustered within schools. In estimating statistical power, classrooms were not explicitly modeled as clustered within schools for two reasons. First, all grade 6 regular classrooms, rather than a random sample, were included in the analytic sample used to estimate the impact of CMP2 on student outcomes (Schochet 2008, p. 2). Second, these classrooms were not conceptualized as representative of a larger population of classrooms within schools (Schochet 2008, p. 22).

To estimate the number of schools needed to achieve statistical power of 0.80, we used results from Schochet (2005, p. 35) table 4, which corresponded with the current study's two-level design and was accompanied by the following assumptions:

- Two-tailed test.
- Equal number of schools randomly assigned to the intervention and control conditions.
- No sampling of class sections within schools.
- Between-school ICC = 0.15.
- Three class sections per school per grade.
- 23 students per class section.
- 80 percent of students in the sample completing both pretest and posttest.
- Proportion of variance explained by school-level covariate with an $R^2 = 0.50$.
- MDES = 0.20.

The MDES of 0.20 standard deviations was considered reasonable based on previous CMP research. The average of the absolute values of the effect sizes reported in Schneider (2000), Riordan and Noyce (2001), and Ridgeway et al. (2003), was 0.24 standard deviations.

The minimum number of schools required to achieve statistical power as specified in this study was 67 (table B1). To guard against the potential loss of schools, three additional schools were recruited to increase the number of schools to be recruited and randomly assigned to conditions to 70.

Table B1. Required school sample sizes to detect target effect sizes for a school randomized design with school-level clustering only

MDES	Schools required to detect an impact ($R^2 = 0.5$)
.10	259
.20	67
.25	44
.33	26

Source: Power analysis conducted during study design.

Appendix C. Procedure and Probability of Assignment to Study Conditions

Random assignment procedures

An illustration of how the random assignment procedures were implemented using Microsoft Excel® is provided in table C1. These examples have been generated for descriptive purposes only, and do not represent actual jurisdiction/schools involved in this study.

Jurisdiction A contained an even number of schools; jurisdiction B contained an odd number of schools. Random assignment was conducted in three steps. All random numbers were generated using the random number generator in Microsoft Excel.

Step one: Randomize school order

Schools were listed by jurisdiction, and each was assigned a random number between 0 and 1 (table C1, panel 1). Schools were then sorted within jurisdiction in ascending order by their random number to remove any list effects (panel 2). In subsequent steps, the randomized order of schools in this table remained fixed.

Table C1. School order randomized using Microsoft Excel

Panel 1: Schools assigned a corresponding random number		Panel 2: Schools sorted by ascending random number	
Column A	**Column B**	**Column B**	**Column A**
Jurisdiction 3 (two schools)		**Jurisdiction 3 (two schools)**	
School E	0.036216002	0.028881803	School F
School F	0.028881803	0.036216002	School E
Jurisdiction 4 (three schools)		**Jurisdiction 4 (three schools)**	
School G	0.887141781	0.760390071	School I
School H	0.939366508	0.887141781	School G
School I	0.760390071	0.939366508	School H

Source: Study records.

Step two: Randomize condition labels

The two available conditions (CMP2/intervention and control) were listed alternately, beginning with CMP2 (table C2, panel 1, column C). In jurisdictions with an odd number of schools, the last row in the list was given an additional condition label to balance the odds of assignment to a particular condition (panel 1, jurisdiction 4).

Within jurisdiction, each condition label was assigned a random number between 0 and 1 (panel 1, columns C and D). The labels (but not schools) were then sorted by this new ascending random number (panel 2).

Table C2. Order of condition labels randomized using Microsoft Excel™

Panel 1: Condition labels assigned a corresponding random number		Panel 2: Condition labels sorted by ascending random number	
Column C	**Column D**	**Column D**	**Column C**
Jurisdiction 3 (two schools)			**Jurisdiction 3 (two schools)**
CMP2	0.08783139	0.03437994	Control
Control	0.03437994	0.08783139	CMP2
Jurisdiction 4 (three schools)[a]			**Jurisdiction 4 (three schools)[a]**
CMP2	0.39000269	0.36224626	Control
Control	0.36224626	0.39000269	CMP2
CMP2	0.42203782	0.39989197	Control
Control	0.39989197	0.42203782	CMP2

a. Jurisdictions with an odd number of schools were provided with an extra condition label to balance the odds of assignment to a particular condition, as shown in this example.
Source: Study records.

Step three: Assign condition labels to schools

The randomly ordered schools (see table C1, panel 2, column A) and condition labels (see table C2, panel 2, column C) were placed next to each other and used to assign schools to conditions (table C3).

For example, in jurisdiction 3, this step resulted in School F being assigned to the control condition and School E to the intervention condition. A condition label that ended up unassigned for any jurisdiction with an odd number of schools was ignored.

Table C3. Schools assigned to conditions

Schools sorted by ascending random number *from table C-1, Panel 2*	Condition labels sorted by ascending random number *from table C-2, Panel 2*
Column A	Column C
Jurisdiction 3 (two schools)	**Jurisdiction 3 (two schools)**
School F	Control
School E	CMP2
Jurisdiction 4 (three schools)[a]	**Jurisdiction 4 (three schools)[a]**
School I	Control
School G	CMP2
School H	Control
	CMP2[a]

a. The extra condition label was ignored.
Source: Data from randomization files.

Unbalanced allocation

An unbalanced allocation does not threaten the internal validity of the study as long as random assignment was implemented properly. It can, however, affect the precision of the estimate if the unbalanced allocation is extreme (Bloom 2005). The two-school difference favoring the intervention group in this study (36 intervention schools and 34 control schools) is too small to affect the statistical precision of the CMP2 impact estimate (Bloom 2005, p. 134).

Appendix D. Student Math Interest Inventory

Student ID: Please write the number in the boxes, and fill in the appropriate circles below.

	☐	☐	☐	☐	☐	☐	☐	☐	☐
0	○	○	○	○	○	○	○	○	○
1	○	○	○	○	○	○	○	○	○
2	○	○	○	○	○	○	○	○	○
3	○	○	○	○	○	○	○	○	○
4	○	○	○	○	○	○	○	○	○
5	○	○	○	○	○	○	○	○	○
6	○	○	○	○	○	○	○	○	○
7	○	○	○	○	○	○	○	○	○
8	○	○	○	○	○	○	○	○	○
9	○	○	○	○	○	○	○	○	○

Student Name (Please Print):_________________________________ **Gender (Circle One):** ☐Male ☐ Female

Teacher:_______________________________________ **School**: ____

Student Math Interest Inventory

Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

Directions

We are trying to understand what students think about the work they do for mathematics class. On the following pages are some examples of what students might think. Please give us your rating for each question.

Different students have different interests, so there are no right or wrong answers. Your answers will **not** be used toward your grade and your teacher will not look at your answers. Please answer these questions honestly, and tell us what you **really** think.

Please bubble in the choice that best describes what you think. You can use a pen or a pencil. If you make a mistake, either erase it or cross it out and completely bubble the correct choice.

<u>Practice Question</u>

How good at science are you?

Not at all Good ☐☐☐☐☐☐☐ Very Good

*If you are OK at science check the middle box. If you good at science check one of the 3 boxes to the right, only check the far right box if you are **very good** at science. If you are not so good at science then check one of the 3 boxes to the left, only checking the far left box if you think you are **not at all good** at science.*

1. In general, I find working on math assignments

Very Boring ☐☐☐☐☐☐☐ Very Interesting

2. How much do you like doing math?

Not Very Much ☐☐☐☐☐☐☐ Very Much

3. Is the amount of effort it will take to do well in advanced high school math courses worthwhile to you?

Not Very Worthwhile ☐☐☐☐☐☐☐ Very Worthwhile

4. I feel that, to me, being good at solving problems which involve math or reasoning mathematically is

Not at all Important ☐☐☐☐☐☐☐ Very Important

5. How important is it to you to get good grades in math?

Not at all Important ☐☐☐☐☐☐☐ Very Important

6. How useful is learning school math for what you want to do after you graduate from high school or college and go to work?

Not Very Useful ☐☐☐☐☐☐☐ Very Useful

7. How useful is what you learn in school math for your daily life outside school?

Not at all Useful ☐☐☐☐☐☐☐ Very Useful

8. Compared to other students, how well do you expect to do in math this year?

Much Worse Than Other Students ☐☐☐☐☐☐☐ Much Better Than Other Students

9. How well do you think you will do in your math course this year?

Very Poorly ☐☐☐☐☐☐☐ Very Well

10. How good at math are you?

Not at all Good ☐☐☐☐☐☐☐ Very Good

11. If you were to order all the students in your math class from the worst to the best in math, where would you put yourself?

The Worst ☐☐☐☐☐☐☐ The Best

12. How have you been doing in math this year?

Very Poorly ☐☐☐☐☐☐☐ Very Well

13. In general, how hard is math for you?

Very Easy ☐☐☐☐☐☐☐ Very Hard

14. Compared to most other students in your class, how hard is math for you?

Much Easier ☐☐☐☐☐☐☐ Much Harder

My Easiest Course	□□□□□□□	My Hardest Course
Not Very Hard	□□□□□□□	Very Hard
A Little	□□□□□□□	A Lot
A Little	□□□□□□□	A Lot
Much Harder in Math than in other Subjects	□□□□□□□	Much Harder in other Subjects than in Math

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless such collection displays a valid OMB control number. The valid OMB control number for this information collection is 1850-0834. *The time required to complete this information collection is estimated to average 5 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, D.C. 20202-4651. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: Rafael Valdivieso, Institute of Education Science, 555 New Jersey Ave, NW, Room 506E, Washington, D.C. 20208-550*

Confirmatory factor analysis for PTV factor

Eccles and Wigfield (1995) uncovered the PTV factor through an exploratory factor analysis and confirmed its structure through a confirmatory factor analysis. However, their confirmatory factor analysis was conducted on a student sample that comprised White adolescents in grades 5–12. The student sample for the current study is diverse in race/ethnicity and restricted to grade 6. Therefore, before using student scores for the PTV domain as a secondary outcome in the impact analysis, a confirmatory factor analysis was conducted using the CMP2 pretest sample.

The PTV data fit the three-factor confirmatory factor analysis model. Although the chi-squared test of exact model fit was statistically significant due to the large sample size, other fit indices indicated a good fit. Comparative fit index and Tucker-Lewis index values greater than 0.95 suggested a good model fit (Hu and Bentler 1999), as did weighted root mean square residual values less than or close to 1 (Yu 2002). All standardized factor loadings were high (> 0.60) and statistically significantly different from zero. Estimated interfactor correlations were moderate to high. Cronbach's alpha was adequately high for Interest but rather low for Importance and Utility. Cronbach's alpha for the seven-item scale was 0.76.

Based on these results, the factor structure confirmed in the Eccles and Wigfield (1995) sample was also confirmed in the CMP2 pretest sample. The sample characteristics, standardized parameter estimates and coefficient alphas for the factors are in table D1

The data fit the three-factor confirmatory factor analysis model—and the equivalent hierarchical confirmatory factor analysis model in which the three first-order factors loaded on a single second-order factor—very well (chi-squared = 44.465, *df* = 11, comparative fit index [CFI] = 0.998, Tucker-Lewis index [TLI] = 0.996, root mean squared error of approximation [RMSEA] = 0.024, weighted root mean square residual [WRMR] = 0.668). Although the chi-squared test of exact model fit was statistically significant due to the large sample size, other fit indices were in the good fit range. CFI and TLI values greater than 0.95 suggested good model fit (Hu and Bentler 1999), as did WRMR values less than 1 or close to 1 (Yu 2002). According to Browne and Cudeck (1993), RMSEA values less than 0.05 indicate close fit. Thus, the factor structure of the seven-item scale was confirmed in the current sample.

Table D1. Parameter estimates for the three-factor confirmatory factor analysis model and coefficient alpha for the factors

	Interest	Importance	Utility
Standardized factor loadings			
Q1	.858*		
Q2	.813*		
Q3		.621*	
Q4		.720*	
Q5		.636*	
Q6			.601*
Q7			.693*
Interfactor correlations			
Interest	1		
Importance	.635*	1	
Utility	.566*	.806*	1
Coefficient alpha	.79	.57	.53

* Estimates statistically significantly different from zero.
Source: Confirmatory factor analysis conducted by the study team using student pretest PTV data.

To test factorial invariance of the above three-factor model across gender and experimental groups, factor loadings, factor variances, and interfactor covariances were constrained to be equal across groups. The three-factor structure appeared to be invariant across gender (chi-squared = 113.522 [73.474 contributed by females, *n* = 2,472; and 40.047 contributed by males, *n* = 2,424], *df* = 64, CFI = 0.997, TLI = 0.998, RMSEA = 0.018, WRMR = 1.25) and study condition (chi-squared = 98.008 [63.821 contributed by the experimental group, *n* = 2,563; and 34.187 contributed by the control group, *n* = 2,578], *df* = 64, CFI = 0.998, TLI = 0.999, RMSEA = 0.014, WRMR = 1.144).

Appendix E. Teacher Surveys

Exhibit. Background

Thank you for participating in this survey. All of your personal information will not be disclosed. Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district, school, or teacher. We will not provide information that identifies you or your district to anyone outside the study team, except as required by law.

1. Name (First & Last): ______________________________
2. School: ____
3. Date of birth (MM/DD/YY)___/___/____
4. Gender (check one) ☐ Male ☐ Female
5. Which of the following best describes your race/ethnicity? *(check all that apply)*

   ☐ White/Caucasian ☐ Asian/Asian ☐ American Hispanic/Latino

   ☐ Black/African-American ☐ American Indian ☐ Other
6. Years working in current school: _______

   Years working in current district: ____Total years teaching experience: ________
7. Highest degree (only check one)

   ☐ Bachelors Major:______________ ☐ Masters Major:________________

   ☐ Ph.D Major:________________

   Number of upper division college mathematics courses taken (above Calculus):

   ____
8. Hours of professional development in mathematics in last 3 years _______
9. Approximate number of 6^{th} grade students you will teach this year _________
10. Have you ever used *Connected Mathematics Project* curriculum before?

    ☐ Yes ☐ No If Yes, for how long? _______Yr(s)._______

Exhibit. Monthly online survey[49]

Thank you for taking the time to fill out this monthly report. Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual.

Please thoroughly answer all questions and provide examples as needed.

1. What unit and associated investigations did you complete this month? (Check all that apply). **Note: Your responses for previous months are checked below to show all of the units you have completed so far.** Please add the additional books completed this month by checking the box in front of the title of the book.

   ☐ Prime Time ☐ Bits and Pieces I ☐ Bits and Pieces II

   ☐ Bits and Pieces III ☐ Shapes and Designs ☐ Data About Us

   ☐ How Likely Is It ☐ Covering and Surrounding

2. Thinking back on the past month, what went particularly well in the unit/investigations that you would recommend others to use/do?

3. Did you face any difficulties with the unit/investigations you taught last month? If so, please describe the difficulties and how you handled them.

4. Approximately how much time did you and your students spend on Connected Mathematics 2 activities each week during the past month? (Choose one response that best approximates the time you spend weekly).

   ☐ Less than 1 hour ☐ 1-2 hours ☐ 2-3 hours

   ☐ 3-4 hours ☐ 4-5 hours ☐ 5-6 hours

   ☐ 6-7 hours ☐ 7-8 hours ☐ More than 8 hours

[49] During the implementation year, the monthly online survey included only the first three questions. The last two questions were added for the impact year.

5. What other non-CMP activities did you and your students spend time on during math class this past month? (check all that apply)

☐ State test preparation materials ☐ Math skills and procedures supplements

☐ Below grade level materials ☐ Teacher created curriculum

☐ District created curriculum ☐ School created curriculum

☐ Other published math curriculum ☐ Test taking strategies

Exhibit. End-of-year summary

Thank you for taking the time to fill out this end of the year summary. This survey will be in lieu of the May Monthly Online Survey. Responses to this data collection will be used only for statistical purposes. The reports prepared for this study will summarize findings across the sample and will not associate responses with a specific district or individual.

Please thoroughly answer all questions and provide examples as needed.

1. What unit and associated investigations did you complete this **year**? (Check all that apply).

☐ Prime Time ☐ Bits and Pieces I ☐ Bits and Pieces II

☐ Bits and Pieces III ☐ Shapes and Designs ☐ Data About Us

☐ How Likely Is It ☐ Covering and Surrounding

2. Approximately how much time did you and your students spend on Connected Mathematics 2 activities each week **on average this year**? (Choose one response that best approximates the time you spend weekly).

☐ Less than 1 hour ☐ 1-2 hours ☐ 2-3 hours

☐ 3-4 hours ☐ 4-5 hours ☐ 5-6 hours

☐ 6-7 hours ☐ 7-8 hours ☐ More than 8 hours

3. What other non-CMP activities did you and your students spend time on during math class **on average this year**? (check all that apply)

☐ State test preparation materials ☐ Math skills and procedures
☐ Below grade level materials ☐ Teacher created curriculum

☐ District created curriculum ☐ School created curriculum

☐ Other published math curriculum ☐ Test taking strategies

According to the Paperwork Reduction Act of 1995, no persons are required to respond to a collection of information unless such collection displays a valid OMB control number. The valid OMB control number for this information collection is 1850-0834. The time required to complete this information collection is estimated to average 5 minutes per response, including the time to review instructions, search existing data resources, gather the data needed, and complete and review the information collection. Information will not be provided that identifies you or your district to anyone outside the study team, except as required by law. If you have any comments concerning the accuracy of the time estimate(s) or suggestions for improving this form, please write to: U.S. Department of Education, Washington, D.C. 20202-4651. If you have comments or concerns regarding the status of your individual submission of this form, write directly to: Rafael Valdivieso, Institute of Education Science, 555 New Jersey Ave, NW, Room 506E, Washington, D.C. 20208-550.

Appendix F. Classroom Observation Data Collection

A comparison of the content in the observation protocols used for classroom observations in intervention and control schools is provided in table F1. Not all the items were used in the implementation analysis. The observation protocol for intervention class sections and control class sections are in tables F2 and F3. A description of the observer training is presented, and the inter-rater reliability based on the outcome of the training is in table F4

Table F1. Parallelism of the content of the intervention and control group protocols

	Protocol item		Response scale	
Item category	**Intervention schools**	**Control schools**	**Intervention schools**	**Control schools**
(1) Classroom time	a. Total minutes in class session	Identical	Interval	Same
	b. Total minutes of mathematics instruction per week	Identical	Interval	Same
(2) Student and classroom	1. Gender	Identical	Interval	Same
	2. Average ability	Identical	Categorical	Same
	3. Disruptive behavior	Identical	Dichotomous	Same
	4. Classroom interruptions	Identical	Dichotomous	Same
	5. Classroom layout (sketch)	Identical	Open	Same
	6. Classroom layout (label)	Identical	Open	Same
(3) Teacher allocation of instructional time and activities	7. Minutes spent on each task	Identical	Interval (for each)	Same
	8. Transition TIME	Identical	Dichotomous	Same
	9. Percent Engaged in Launch, Explore, Summarize	Percent engaged in time 1, time 2, time 3	Interval ratio (for each)	Same
	10. Student problem solving and connections	Student problem solving and connections	Categorical (all that apply)	Same
	11. Inadequate classroom physical features	Inadequate classroom physical features	Dichotomous (all that apply)	Same
(4) Teaching use of curriculum materials	12. Materials used to teach lesson	Similar	Dichotomous (all that apply)	Same
	13. Student in-class activities	Describe student in-class activities	Dichotomous (all that apply)	Open
	14. Non-CMP2 materials	Other technology/materials used:	Open	Same
(5) Teacher pedagogy, content expertise, and efficacy	15. Pedagogy	Pedagogy	Interval (from 1 to 5)	Same

Item category	Protocol item		Response scale	
	Intervention schools	**Control schools**	**Intervention schools**	**Control schools**
	16. Instructional practices	Identical	Dichotomous (for each)	Same
	17. Content expertise	Identical	Interval	Same
	18. Classroom efficacy	Identical	Interval	Same
	19. Vocabulary terms	Identical	Open	Same
(6) Teaching instructional strategies	20. CMP2 book title	Textbook	Open	Same
	21. Prelaunch activities	Describe teaching strategies	Open	Same
	22. Launch	Assessment	Open	Same
	23. Explore	Feedback/grading	Open	Same
	24. Summarize	Notes	Open	Same
	25. Summarize next day	Not applicable	Dichotomous	Not applicable
(7) Teacher assessment and grading	26. Assessment	Not applicable	Dichotomous (all that apply)	Not applicable
	27. Feedback/grading	Not applicable	Dichotomous (all)	Not applicable
Notes	28. Notes	Not applicable	Open	Not applicable

Source: Author analysis of observation protocols.

Table F2. Intervention school observation protocol

Prentice Hall CMP2 Intervention Classroom Observation Protocol

Teacher Name:_______ Period:________ School:____________ Date:______ Observer Initials:_____ Total # min. in class session:______

**Total Minutes of Math Instruction Per Week: _________________ (Based on teacher response

STUDENT AND CLASSROOM VARIABLES	INSTRUCTIONAL VARIABLES	TEACHING MATERIALS
1. Gender: # *Male*______ # *Female*______ **2. Average ability of students in this class period (ask teacher—circle):** Special Needs Below-Level Average Advanced ELL Varies **3. Disruptive student behavior (check only one):** ☐ No significant classroom disruptions ☐ Small disruptions from multiple sources ☐ Small # of students persistently disruptive ☐ Large # of students disruptive throughout ☐ Other behavior issues: ___________________________________ **4. Number of classroom interruptions:** <3 >3 (Only consider an event a disruption if student learning is interrupted.)	**7. Estimate the # of minutes spent on each of the following tasks:** a. ____ Routines (e.g., checking homework, taking role) b.____ Teacher-directed lecture c.____ Class discussion d.____ Small group activities e.____ Paired student activities f.____ Student independent work time g.____ Homework review h.____ Test/Chapter/unit review i.____ In-class quiz or test j.____ State standardized testing k.____ School activities (e.g., announcements, assemblies) l.____ Student discipline/interruptions m.____ Other: _____________________________ **8. Transition time:** <5min >5min	**12. Check off each of the materials used to teach the lesson:** a. ☐ Student Edition textbook b. ☐ Teacher Edition textbook c. ☐ Special Needs Handbook d. ☐ Spanish Resources e. ☐ Lab Sheets f. ☐ CMP2 transparencies g. ☐ ELMO (document projector) h. ☐ Teacher-created transparencies i. ☐ Manipulatives j. ☐ Vocabulary k. ☐ Teacher Express CD-ROM l. ☐ Exam View CD-ROM m. ☐ None of the above n. ☐ Other: ____________________

Note: Items 5–6, 9–11, and 13–14 are on the following page.

STUDENT AND CLASSROOM VARIABLES	INSTRUCTIONAL VARIABLES	TEACHING MATERIALS

STUDENT AND CLASSROOM VARIABLES

5. and 6. Sketch and label features of the classroom layout:

INSTRUCTIONAL VARIABLES

9. Estimate the percent of students engaged in the lesson during each phase of L, E, S:

LAUNCH ____ %
EXPLORE ____ %
SUMMARIZE ____ %

10. Did students do the following?	**Class discussion**	**Group Work**	**Pair Work**
Answer each other's questions			
Make connections to previous lessons			
Introduce more than one way to approach a problem			
Take turns answering teacher probes (not one/few students dominating the discussion)			
Collaborate to solve a problem			

Note: *Please Mark Yes, No, or N/A for each box above. Use N/A if students did not work in that kind of arrangement (i.e. pairs)*

11. Circle any physical features that are NOT adequate:
Lighting   Outside/inside noise   Space for chairs/desks   Temperature

TEACHING MATERIALS

13. Check off each of the activities students worked on during class:

a. ☐ ACE problems
b. ☐ Mathematical Reflections
c. ☐ Exercises (A, B, C, D, etc.)
d. ☐ Partner Quizzes
e. ☐ Check-ups
f. ☐ Self-Assessments
g. ☐ Question Bank Problems
h. ☐ Multiple Choice Problems
i. ☐ Notebook Checklists
j. ☐ Other: ______________________
k. ☐ Homework Assigned:

14. Non-CMP2 Materials:
____
____
____

TEACHER VARIABLES	DESCRIPTION OF TEACHING STRATEGIES
15. Pedagogy: (mark on continuum) "Sage on the stage" "Guide on the side" ---x-- **16. Did the teacher do any of the following?** a. ☐ Present student learning goals related to the lesson/activity b. ☐ Explain rules and definitions c. ☐ Instruct students to look at the textbook while teacher talked about it d. ☐ Allow students to answer each other's questions e. ☐ Allow students to work with/help each other f. ☐ Encourage curiosity and creativity in students g. ☐ Expect students to engage in complex thinking h. ☐ Connect concepts taught in class to things students already know i. ☐ Connect concepts taught in class to the "real world" j. ☐ Establish daily classroom routines (e.g., homework collection, notebook checks, etc.) k. ☐ Use alternative teaching strategies if students fail to understand the lesson l. ☐ Assess students' prior knowledge of a concept m. ☐ Provide positive encouragement to students n. ☐ Make sure all students are on board before moving on o. ☐ Other: ____________________ 17. Overall content expertise: 1 2 3 4 5 18. Overall classroom efficacy: 1 2 3 4 5 19. Vocabulary terms referenced:	**20. CMP2 Problem/Unit:** ____________________ **21. Prelaunch Activities:** Teacher setup/explanation prior to launch ____________________ ____________________ ____________________ **22. Launch** (include a brief description of the "real world" applications used to launch the lesson)**:** ____________________ ____________________ ____________________ ____________________ **23. Explore** (include examples of questions the teacher directed to students as well as descriptions of student/teacher and student/student interactions)**:** ____________________ ____________________ ____________________ ____________________ ____________________ ____________________ ____________________ **24. Summarize** (focus on adherence to CMP2 philosophy, noting whether or not the teacher generated his/her summary based on student responses and work)**:** ____________________ ____________________ ____________________ ____________________**25) Is teacher planning to do/complete "Summarize" tomorrow? Yes/No**

ASSESSMENT, FEEDBACK, and GRADING

26. Assessment: (Check all methods of assessment that were observed):
☐ Warm-up ☐ Individual Quiz ☐ Unit Test ☐ Unit Project ☐ Whiteboards or student response system

27. Feedback/Grading: (Check methods of feedback, grading, and opportunities for revision that are observed):
☐ Collection of student work to be graded ☐ Graded work returned to students ☐ Feedback written on student work in the form of a grade, points, etc.
☐ Feedback written on student work in the form of comments ☐Opportunities for students to revise work ☐ Rubric used in grading

Notes:____________________
____

Table F3. Control school observation protocol

Prentice Hall CMP2 Control Classroom Observation Protocol

Teacher Name:_________ Period:________ School:_______ State:_____ Date:______ Observer Initials:___ Total # min. in class session:______

**Total Minutes of Math Instruction Per Week: ________________ (Based on teacher response)

STUDENT AND CLASSROOM VARIABLES	INSTRUCTIONAL VARIABLES	TEACHING MATERIALS
1. Gender: # *Male* ___ # *Female* ______ **2. Average ability of students in this class period (ask teacher—circle):** Below-Level Average Special Needs Advanced ELL Varies **3. Disruptive student behavior (check only one):** ☐ No significant classroom disruptions ☐ Small disruptions from multiple sources ☐ Small # of students persistently disruptive ☐ Large # of students disruptive throughout ☐ Other behavior issues: ______________________________ **4. Number of classroom interruptions:** <3 >3 (Only consider an event a disruption if student learning is interrupted.)	**7. Estimate the # of minutes spent on each of the following tasks:** _Routines (e.g., checking homework, taking role) a. ____ Teacher-directed lecture b. c. ____ Class discussion d. ____ Small group activities e. ____ Paired student activities f. ____ Student independent work time g. ____ Homework review h. ____ Test/Chapter/unit review i. ____ In-class quiz or test j. ____ State standardized testing _School activities (e.g., announcements) k. l. ____ Student discipline/interruptions m. ____ Other: **8. Transition time:** >5min <5min	**12. Check off each of the materials used to teach the lesson:** a. ☐ Textbook b. ☐ Textbook publisher transparencies c. ☐ Teacher-created transparencies d. ☐ Textbook publisher handouts/ worksheets e. ☐ Teacher-created handouts f. ☐ Student workbooks g. ☐ Manipulatives h. ☐ Materials/problems from online resources i. ☐ Other publisher materials j. ☐ Other teacher-created materials k. ☐ None of the above

Note: Items 5–6, 9–11, and 13–14 are shown on the following page.

5. and 6. Sketch and label features of the classroom layout:

9. Estimate the percent of students engaged in the lesson at the indicated times:

TIME 1 (5-10 minutes into class): ____ %
TIME 2 (25-30 minutes into class): ____ %
TIME 3 (10-15 minutes before class ends): ____ %

Note: Write 0% for no engagement, -1 if that part did not occur.

10. Did students do the following?	**Class discussio n**	**Group Work**	**Pair Work**
Answer each other's questions			
Make connections to previous lessons			
Introduce more than one way to approach a problem			
Take turns answering teacher probes (not one/few students dominating the discussion)			
Collaborate to solve a problem			

Note: *Please Mark Yes, No, or N/A for each box above. Use N/A if students did not work in that kind of arrangement (i.e. pairs)*

11. Circle any physical features that are NOT adequate:

Lighting Outside/inside noise Space for chairs/desks Temperature

13. Describe the activities students worked on during class:

____
____
____
____
____
____
____
____
____
____
____
____
____

14. Other technology/materials used:

TEACHER VARIABLES	DESCRIPTION OF TEACHING STRATEGIES
15. Pedagogy: (mark on continuum) "Sage on the stage" "Guide on the side" --x-- **16. Did the teacher do any of the following?** a. ☐ Present student learning goals related to the lesson/activity b ☐ Explain rules and definitions c. ☐ Instruct students to look at the textbook while teacher talked about it d. ☐ Allow students to answer each other's questions e. ☐ Allow students to work with/help each other f. ☐ Encourage curiosity and creativity in students g. ☐ Expect students to engage in complex thinking h. ☐ Connect concepts taught in class to things students know i. ☐ Connect concepts taught in class to the "real world" j. ☐ Establish daily classroom routines (e.g., homework collection, notebook checks, etc.) k. ☐ Use alternative teaching strategies if students fail to understand the lesson l. ☐ Assess students' prior knowledge of a concept m. ☐ Provide positive encouragement to students n. ☐ Make sure all students are on board before moving on o. ☐ Other: 17. Overall content expertise: 1 2 3 4 5 18. Overall classroom efficacy: 1 2 3 4 5 19. Vocabulary terms referenced:	**20. Textbook Problem/Chapter/Unit, p#:** ______________________________ ______________________________ ______________________________ **Teaching strategies used to deliver today's lesson:** **21. Time 1** (First 5-10 min of lesson: brief description of the way the teacher introduced the lesson) ______________________________ ______________________________ **22. Real World Connections** (During the introduction of the lesson, describe any "real world" applications used) ______________________________ ______________________________ **23. Time 2** (25-30 minutes into class: provide examples of questions the teacher directed to students, descriptions of student/teacher and student/student interactions) ______________________________ ______________________________ **24. Time 3** (10-15 minutes before class ends: describe whether and how the teacher summarized the lesson with the students. ______________________________ ______________________________ **25. Is the teacher planning on finishing the lesson tomorrow?** **Yes No**

ASSESSMENT, FEEDBACK, and GRADING

26. Assessment: (Check all methods of assessment that were observed):

☐ Warm-up ☐ Individual Quiz ☐ Unit Test ☐ Unit Project ☐ Whiteboards or student response system

27. Feedback/Grading: (Check methods of feedback, grading, and opportunities for revision that are observed):
☐ Collection of student work to be graded ☐ Graded work returned to students ☐ Feedback written on student work in the form of a grade, points, etc.
☐ Feedback written on student work in the form of comments ☐Opportunities for students to revise work ☐ Rubric used in grading

Notes:____
____
____

Observer training

Observer training was conducted using approximately 45-minute videos presenting CMP2 and non-CMP2 lessons. The CMP2 videos were filmed in CMP2 class sections in schools that were not participating in the study. These videos were borrowed from a university professor who had permission to use them for training. The videos of non-CMP2 lessons were taken from publicly available online materials for PD.

First, two primary raters established benchmark ratings for reliability for each video. For both the CMP2 and non-CMP2 lesson videos, the two primary raters had an overall average inter-rater reliability rate of 96 percent.

Before the implementation year began, when only intervention schools were observed, 11 observers were trained on the classroom observation protocol with the videos for the CMP2 lessons. While viewing the videos, each potential observer rated the classroom session using the observation protocol. To qualify as a study observer, each potential observer was required to meet an 80 percent agreement with the benchmark for reliability. The classroom observers met this standard with an average 84 percent inter-rater agreement with the benchmark ratings.

During the impact year, both intervention and control school classrooms were observed. A second training session was therefore conducted for control school observations using videos of non-CMP2 lessons. All 11 of the original observers had 89 percent agreement with the benchmark ratings for control-site observation protocols.

Five additional observers were trained for the impact year and achieved average ratings of 87 percent on the control videos and 84 percent on the CMP2 videos (table F4).

Table F4. Inter-rater reliability from observation protocol training (percent)

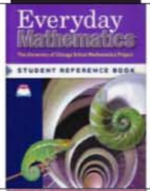
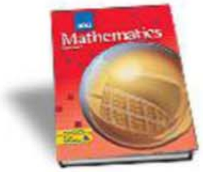
			Site-visit reliability		
Observers	**Participated in classroom observations for the implementation year**	**Participated in classroom observations for the impact year**	**Reliability on videos of teachers using CMP2 but not in the CMP2 study sample**	**Reliability on videos of teachers using curricula other Than CMP2, but not in the study sample**	**Average reliability score**
A	Yes	Yes	80	94	87
B	Yes	Yes	87	89	88
C	Yes	Yes	82	85	84
D	Yes	Yes	81	87	84
E	Yes	Yes	82	83	83
F	Yes	Yes	81	91	86
G	Yes	Yes	88	89	89
H	Yes	Yes	87	97	92
I	Yes	Yes	85	91	88
J	Yes	Yes	83	90	87
K	Yes	Yes	85	88	87
L	No	Yes	83	88	86
M	No	Yes	87	85	86
N	No	Yes	84	88	86
O[a]	No	Yes	80	88	84
P[a]	No	Yes	84	86	85
Average			84	89	86

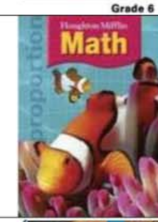
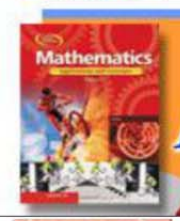
a. Lab personnel new to the study trained in preparation for the impact year.
Source: Analysis of observer training results.

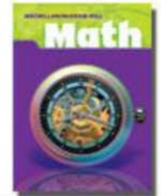
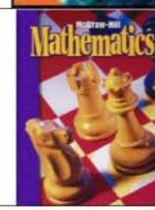
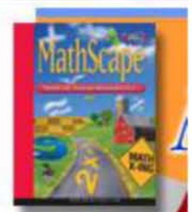
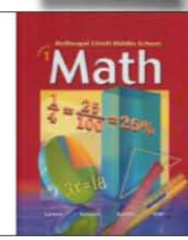
Exhibit. Control school curriculum verification for spring site visits

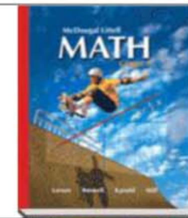
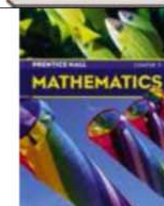
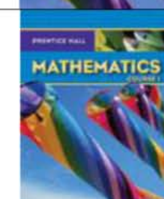
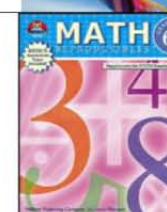
Control School Curriculum Verification for Spring Site Visits

When completing the site visit protocol, please verify the curriculum in use by control teachers and specify the title, publisher, and curriculum number on the paper protocol. Some of the titles are similar, so we have provided an image of the cover of the student book to assist you in verifying the curriculum. If the curriculum in use is not on this list, please provide the title, publisher, and a brief description of the cover of the book so we can make sure to identify it correctly.

Curriculum Number	Title of 6th Grade Mathematics Curriculum	Year	Publisher	Picture of Cover
1	EnVision Math	2008	Scott Foresman- Addison Wesley	
2	Everyday Math	2007 (3rd Edition) Earlier Editions, too	Wright Group/McGraw Hill	
3	Harcourt Math Grade 6	2007	Harcourt Brace	
4	Holt Course 1	2007	Holt, Rinehart, Winston	

Curriculum Number	Title of 6th Grade Mathematics Curriculum	Year	Publisher	Picture of Cover
5	Houghton Mifflin Grade 6 Math	2005	Houghton Mifflin	
6	Houghton Mifflin Grade 6 Math	2007	McDougal-Litel/Houghton Mifflin	
7	HSP Math Grade 6	2009	Houghton Mifflin-Harcourt	
8	Mathematics Applications and Concepts Course 1	2004	Glencoe	
9	Math Connects	2009	Glencoe MacMillan McGraw Hill	

Curriculum Number	Title of 6th Grade Mathematics Curriculum	Year	Publisher	Picture of Cover
10	Math	2003	MacMillan McGraw Hill	
11	Math Triumphs Grade 6	2008	MacMillan McGraw Hill	
12	Mathematics Grade 6	2002	McGraw Hill Mathematics	
13	Mathscape Course 1	2005	McGraw-Hill Glencoe	
14	Math Course 1	2002	McDougall Littell	

Curriculum Number	Title of 6th Grade Mathematics Curriculum	Year	Publisher	Picture of Cover
15	Math Course 1	2007	McDougall Littell	
16	Prentice Hall Core Math Course 1	2004	Prentice Hall	
17	Prentice Hall Core Math Course 1	2008	Prentice Hall	
18	Properties of Whole Numbers Milliken Publishing	unknown	Milliken Publishing	
19	Saxon Math Course 1	2006	Saxon	

Appendix G. Equations to Estimate the Impact of CMP2

When the main hierarchical linear models (HLM) and the alternative models used in the sensitivity analyses were estimated, a pretest at both levels in the two-level model were included—for three reasons. First, we did not want to assume that the within-school pretest relationship with the outcome is the same as the between-school pretest relationship with the outcome (Raudenbush and Bryk 2002). Second, adding a pretest at level 1 as a predictor reduces within-school variance. Third, adding a pretest at both levels 1 and 2 is recommended when analyzing data generated by a cluster RCT design, according to Burghardt et al. (2009, p. 8).

In estimating these models, the variables were centered. Centering changes the interpretation of the coefficients. The pretest was school-mean centered at level 1 because the level-1 pretest coefficient was to represent the pooled within-school student pretest relationship (as opposed to a mix of between and within relationships if the pretest were grand-mean centered). The pretest was grand-mean centered at level 2 because the level two school-mean pretest coefficient was to represent the between-school pretest relationship with the outcome.

Furthermore, all the level-2 covariates were grand-mean centered so that the intercept represented the adjusted control group mean and that the effects of a level-2 independent variable could be evaluated (CMP2) while controlling for other level-2 covariates. Grand-mean centering level-2 covariates in a two-level model is a common practice, for example, to make the interpretation of the intercept meaningful (O'Connell and McCoach 2008).

To conclude, for the two-level models described in the next section, there was a very specific purpose for including the pretest at both levels of the two-level model and for using school-mean centering at level 1 and grand-mean centering at level 2. This purpose also applied to the one three-level model but with group-mean centering of the level-1 and level-2 variables and grand-mean centering of level-3 variables.

Main model

Level 1 (student) equation:

$$Y_{ij} = \beta_{0j} + \beta_{1j}*(\text{PretestStdt})_{ij} + \beta_{2j}*(\text{PretestStdtMiss})_{ij} + r_{ij}.$$

Level 1 variables:

Y_{ij} – outcome for student *i* in school *j*.

$(\text{PretestStdt})_{ij}$ – pretest score for student *i* in school *j*, school-mean centered, with an imputed constant (the grand mean of student pretest scores for the full sample, irrespective of group membership) when the score is missing.

$(\text{PretestStdtMiss})_{ij}$ – the indicator variable for missing pretest score for student *i* in school *j*, school mean centered, scored as 1 when the student's pretest score is missing and as 0 when student's pretest score is observed.

Level 1 coefficients:

β_{0j} – average outcome of students in school *j* adjusted for proportion of students missing pretest scores.
β_{1j} – expected increase in TerraNova score for every unit increase of student pretest score for school *j* controlling for proportion of students missing pretest scores.
β_{2j} – average difference in TerraNova score between students who missed and did not miss pretest adjusted for student pretest score for school *j*.
r_{ij} – random error associated with student *i* in school *j*; $r_{ij} \sim N(0, \sigma^2)$.

The school average outcome estimated by the level-1 intercept β_{0j} was modeled as a function of the intervention (CMP2) at level 2, the school level, controlling for the school average pretest scores on the TerraNova mathematics subtest. Further, even though intervention and control groups were formed using random assignment, there was always a chance that a particular sample might have a statistically significant difference on some measured baseline characteristic.[50] To control for such chance differences, any baseline variables for which there was a statistically significance difference between intervention and control schools at $p < .05$ were included in the model at level 2. Thus, level 2 was specified as follows:

Level 2 (school) equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\text{CMP2})_j + \gamma_{02}*(\text{PretestSch})_j + \gamma_{03}*(\text{Juris1})_j + \gamma_{04}*(\text{Juris2})_j + \gamma_{05}*(\text{Juris3})_j + \gamma_{06}*(\text{SchoolUrban})_j + \gamma_{07}*(\text{TeacherWhite})_j + \gamma_{08}*(\text{TeacherMathMajor})_j + \gamma_{09}*(\text{StudentBlack})_j + \gamma_{010}*(\text{StudentWhite})_j + u_{0j}.$$

$$\beta_{1j} = \gamma_{10}.$$

$$\beta_{2j} = \gamma_{20}.$$

Level 2 variables:

$(\text{CMP2})_j$ – an intervention indicator that takes a value of 1 for an intervention school and 0 for a control school.
$(\text{PretestSch})_j$ – average pretest score for students in school *j*, grand-mean centered.
$(\text{Juris1})_j$ – a jurisdiction indicator variable that takes a value of 1 for jurisdiction 1 and 0 for jurisdictions 2, 3, and 4, grand-mean centered.
$(\text{Juris2})_j$ – a jurisdiction indicator variable that takes a value of 1 for jurisdiction 2 and 0 for jurisdictions 1, 3, and 4, grand-mean centered.
$(\text{Juris3})_j$ – a jurisdiction indicator variable that takes a value of 1 for jurisdiction 3 and 0 for jurisdictions 1, 2, and 4, grand-mean centered.
$(\text{SchoolUrban})_j$ – a school locale variable that takes a value of 1 for urban and 0 for rural, suburban, and small city, grand-mean centered.
$(\text{TeacherWhite})_j$ – the percentage of teachers in school *j* that are White, grand-mean centered.

[50] The inclusion of a pretest covariate typically yields improved statistical precision of the parameter estimates (Bloom, Richburg-Hayes, and Black 2007; Raudenbush, Martinez, and Spybrook 2005).

$(\text{TeacherMathMajor})_j$ – the percentage of teachers in school j that majored in Mathematics during college, grand-mean centered.
$(\text{StudentBlack})_j$ – the percentage of students in school j that are Black, grand-mean centered.
$(\text{StudentWhite})_j$ – the percentage of students in school j that are White, grand-mean centered.

Level 2 coefficients:

γ_{00} – adjusted average student outcome across all control schools.
γ_{01} – adjusted average difference in student outcome between the intervention schools and the control schools (the intervention effect).
γ_{02} – adjusted increase in school average TerraNova score for every unit increase in school average pretest score.
γ_{03} – adjusted average difference in student outcome between jurisdiction 1 and jurisdiction 4.
γ_{04} – adjusted average difference in student outcome between jurisdiction 2 and jurisdiction 4.
γ_{05} – adjusted average difference in student outcome between jurisdiction 3 and jurisdiction 4.
γ_{06} – adjusted average difference in student outcome between urban and non-urban schools.
γ_{07} – adjusted increase in school average TerraNova score for every unit increase in the percentage of teachers within school who are White.
γ_{08} – adjusted increase in school average TerraNova score for every unit increase in the percentage of teachers within school who majored in mathematics during college.
γ_{09} – adjusted increase in school average TerraNova score for every unit increase in the percentage of students within school who are Black.
γ_{010} – adjusted increase in school average TerraNova score for every unit increase in the percentage of students within school who are White.
u_{0j} – random error associated with school j on school average outcome where $u_{0j} \sim N(0, \tau_{00})$.
γ_{10} – adjusted average relationship between student pretest score and TerraNova score across schools.
γ_{20} – adjusted average difference across schools between students who missed the pretest and those who did not in TerraNova score.

Of primary interest among the level-2 coefficients was γ_{01}, which represents CMP2's adjusted effect on the student outcome. A statistically significant value of γ_{01} would indicate that CMP2 has an effect on mathematics achievement after adjusting for pretest, jurisdiction, and other baseline covariates for which there was a statistically significant difference between intervention and control schools at $p < .05$. HLM 6.0 software was used to analyze all the HLMs with the default restricted maximum likelihood estimation for the two-level models. To address the secondary research question, we replaced the

TerraNova with the PTV as the outcome (Y_{ij}*) in the model just described and re-estimated the model using the analytic sample for the PTV.

The WWC handbook recommends that WWC reviewers use the square root of the unadjusted, pooled student level variance, but not from the HLM, and not for the cluster level (p. 45). CMP has been reviewed by the WWC and this study could be reviewed as part of an update. Therefore, the effect size was calculated using Hedges' g, consistent with the guidance in appendix B (p. 45) of the WWC Procedures and Standards Handbook (version 2.1).

Models for sensitivity analyses

Unadjusted CMP2 impact

Level 1 (student) equation:

$Y_{ij} = \beta_{0j} + r_{ij}$.

Level 1 variables:

Y_{ij} – outcome for student *i* in school *j*.

Level 1 coefficients:

β_{0j} – average unadjusted outcome of students in school *j*.
r_{ij} – random error associated with student *i* in school *j;* $r_{ij} \sim N(0, \sigma^2)$.

The school average outcome estimated by the level-1 intercept β_{0j} was modeled as a function of the intervention at level 2, the school level. Thus, level 2 was specified as follows:

Level 2 (school) equation:

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(CMP2)_j + u_{0j}$.

Level 2 variables:

$(CMP2)_j$ –an intervention indicator that takes a value of 1 for an intervention school and 0 for a control school.

Level 2 coefficients:

γ_{00} – unadjusted average student outcome across all control schools.
γ_{01} – unadjusted average difference in student outcome between the intervention schools and the control schools (the intervention effect).
u_{0j} – random error associated with school *j* on school average outcome, where $u_{0j} \sim N(0, \tau_{00})$.

Of primary interest among the level-2 coefficients was γ_{01}, which represents CMP2's unadjusted effect on the student. A statistically significant value of γ_{01} would indicate that CMP2 has an effect on mathematics achievement without adjustments for covariates.

CMP2 impact adjusted for urban locale

Level 1 (student) equation:

$Y_{ij} = \beta_{0j} + r_{ij}$.

Level 1 (student) variables:

Y_{ij} – outcome for student *i* in school *j*.

Level 1 coefficients:

β_{0j} – unadjusted average outcome of students in school *j*.
r_{ij} – random error associated with student *i* in school *j*; $r_{ij} \sim N(0, \sigma^2)$.

The school average outcome estimated by the level-1 intercept β_{0j} was modeled as a function of the intervention at level 2, the school level, controlling for urban locale. We controlled for urban locale to analyze how sensitive the unadjusted CMP2 impact estimate was to a statistical control for urban locale. Thus, level 2 (school) was specified as follows:

Level 2 (school) equation:

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\text{CMP2})_j + \gamma_{02}*(\text{SchoolUrban})_j + u_{0j}$.

Level 2 variables:

$(\text{CMP2})_j$ – an intervention indicator that takes a value of 1 for an intervention school and 0 for a control school.
$(\text{SchoolUrban})_j$ – a school locale variable that takes a value of 1 for urban and 0 for non-urban, grand-mean centered.

Level 2 coefficients:

γ_{00} – adjusted average student outcome across all control schools.
γ_{01} – adjusted average difference in student outcome between the intervention schools and the control schools (the intervention effect).
γ_{02} – adjusted average difference in student outcome between urban and non-urban schools.
u_{0j} – random error associated with school *j* on school average outcome where $u_{0j} \sim N(0, \tau_{00})$.

Of primary interest among the level-2 coefficients was γ_{01}, which represents CMP2's main effect on the student outcome adjusted for urban locale. A statistically significant value of γ_{01} would indicate that CMP2 has an effect on mathematics achievement after adjusting for urban locale.

Handling missing data on the pretest

In the main analysis, we included all students who had a valid posttest score in the sample. Students missing a pretest score were assigned a grand-mean pretest score and had a missing pretest indicator set to 1. To test how sensitive the CMP2 impact estimate was to this choice, the models specified earlier were re-estimated using a sample of students who had observed pretest and posttest scores (also known as listwise or case deletion). The model was adjusted by excluding the missing pretest indicator from level 1.

The model specification was as follows.

Level 1 (student) equation:

$Y_{ij} = \beta_{0j} + \beta_{1j}*(\text{PretestStdt})_{ij} + r_{ij}$.

Level 1 variables:

Y_{ij} – outcome for student *i* in school *j*.
$(\text{PretestStdt})_{ij}$ – the pretest score for student *i* in school *j*, school-mean centered.

Level 1 coefficients:

β_{0j} – average outcome of students in school *j*.
β_{1j} – expected increase in TerraNova posttest score for every unit increase of student pretest score in school *j*.
r_{ij} – random error associated with student *i* in school *j*, and $r_{ij} \sim N(0, \sigma^2)$.

Level 2 (school) equation:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\text{CMP2})_j + \gamma_{02}*(\text{PretestSch})_j + \gamma_{03}*(\text{Juris1})_j + \gamma_{04}*(\text{Juris2})_j + \gamma_{05}*(\text{Juris3})_j + \gamma_{06}*(\text{SchoolUrban})_j + \gamma_{07}*(\text{TeacherWhite})_j + \gamma_{08}*(\text{TeacherMathMajor})_j + \gamma_{09}*(\text{StudentBlack})_j + \gamma_{010}*(\text{StudentWhite})_j + u_{0j}.$$

$\beta_{1j} = \gamma_{10}$.

The level-2 variable definitions, as well as centering and interpretation of level-2 coefficients in the main model, apply here. Finally, the adjusted relationship between pretest score and the outcome of student *i* in school *j* (β_{1j}) was modeled as fixed at level 2.

Of primary interest among the level-2 coefficients was γ_{01}, which represents CMP2's main effect on the student outcome. A statistically significant value of γ_{01} would be reason to reject the null hypothesis of no difference between intervention and control

schools in favor of the alternative hypothesis that there was a difference between intervention and control schools. The value of this parameter estimate was compared with the same parameter estimate generated by the corresponding main HLM that used the dummy variable adjustment for missing data on the pretest. The two estimates were compared to determine how sensitive the estimate was relative to the estimate from the main HLM.

Controlling for group differences on covariates at p < .10

Any baseline variables that were not statistically significant between intervention and control groups at $p < .05$ but were statistically significant at $p < .10$ were included in the HLM as a sensitivity analysis. Specifically, each variable was included as a school-level covariate (grand-mean centered) in addition to the pretest school mean covariate (grand-mean centered) to address the confirmatory research question. This analysis indicated whether the estimate and statistical significance were sensitive to excluding these variables from the model. The model specification for this sensitivity analysis mirrors the specification for the main HLM presented earlier in this appendix but with the addition of baseline variables that were statistically significantly different between intervention and control groups at $p < .10$.

Level 1 (student) equation:

$Y_{ij} = \beta_{0j} + \beta_{1j}*(\text{PretestStdt})_{ij} + \beta_{2j}*(\text{PretestStdtMiss})_{ij} + r_{ij}.$

Level 2 (school) equation:

$\beta_{0j} = \gamma_{00} + \gamma_{01}*(\text{CMP2})_j + \gamma_{02}*(\text{PretestSch})_j + \gamma_{03}*(\text{Juris1})_j + \gamma_{04}*(\text{Juris2})_j + \gamma_{05}*(\text{Juris3})_j + \gamma_{06}*(\text{SchoolUrban})_j + \gamma_{07}*(\text{TeacherWhite})_j + \gamma_{08}\ (\text{TeacherMathMajor})_j + \gamma_{09}*(\text{StudentBlack})_j + \gamma_{010}*(\text{StudentWhite})_j + \gamma_{011}*(\text{TeacherBlack})_j + \gamma_{012}*(\text{StudentFemale})_j + u_{0j}.$
$\beta_{1j} = \gamma_{10}.$
$\beta_{2j} = \gamma_{20}.$

Additional baseline variables at level 2:

$(\text{TeacherBlack})_j$ – percentage of teachers in school j that are Black, grand-mean centered.
$(\text{StudentFemale})_j$ – percentage of students in school j that are female, grand-mean centered.

Three-level instead of two-level model

To investigate how sensitive the impact estimate, standard errors, and statistical tests were to the decision to ignore clustering at the classroom level, a three-level model that included the class section at level 2, students at level 1, and schools at level 3 was also estimated. The purpose of this sensitivity analysis was to determine whether estimates of CMP2's standard errors were unaffected. The model specification was as follows:

Level 1 (student) equation:

$Y_{ijk} = \pi_{0jk} + \pi_{1jk}$*(PretestStdt)$_{ijk}$ + π_{2jk}*(PretestStdtMiss)$_{ijk}$ + e_{ijk}.

Level 1 variables:

Y_{ijk} – student outcome for student *i*, in class section *j*, in school *k*.
(PretestStdt)$_{ijk}$ – the pretest score for student *i* in class section *j*, in school *k*, class-mean centered.
(PretestStdtMiss)$_{ijk}$ – indicator variable for missing pretest score for student *i* in class section *j* in school *k*, where 1 = missing pretest score and 0 = did no miss pretest score, grand-mean centered.

Level 1 coefficients:

π_{0jk} – average outcome of students in class section *j*, in school *k* adjusted for proportion of students missing pretest score.
π_{1jk} – adjusted relationship between pretest score and the student outcome in class section *j* in school *k*.
π_{2jk} – expected difference in the outcome measure between students who missed pretest and those who did not in class section *j* in school *k*.
e_{ijk} – random error associated with student *i*, in class section *j*, in school *k* with $e_{ijk} \sim N(0,\sigma^2)$.

The class section average outcome estimated from the above model (level-1 intercept π_{0jk}) was modeled as varying randomly across class sections within schools at level 2. The adjusted relationship between pretest score and the outcome (π_{1jk}) and the difference between missing and nonmissing pretest score (π_{2jk}) was modeled as fixed across class sections within schools. The level-2 specification is as follows.

Level 2 (class section) equation:

$\pi_{0jk} = \beta_{00k} + \beta_{01k}$*(PretestClass)$_{jk}$ + r_{0jk}.
$\pi_{1jk} = \beta_{10k}$.
$\pi_{2jk} = \beta_{20k}$.

Level 2 variables:

(PretestClass)$_{jk}$ – pretest score for class section *j*, in school *k*, school mean centered.

Level 2 coefficients:

β_{00k} – average class section outcome in school *k*.
β_{01k} – relationship between class average pretest score and class average outcome in school *k*.

β_{10k} – average adjusted relationship between student pretest score and student outcome across all class sections, in school *k.*

β_{20k} – average expected difference between missing and non-missing pretest score across all class sections in school *k*.

r_{0jk} – a random error associated with class section *j* in school *k* on class section average student outcome with $r_{0jk} \sim N(0, \tau_\pi)$.

Level 3 (school) equation:

$\beta_{00k} = \gamma_{000} + \gamma_{001}(\text{CMP2})_k + \gamma_{002}(\text{PretestSch})_k + \gamma_{003}(\text{Juris1})_k + \gamma_{004}(\text{Juris2})_k + \gamma_{005}(\text{Juris3})_k + \gamma_{006}{*}(\text{SchoolUrban})_k + \gamma_{007}{*}(\text{TeacherWhite})_k + \gamma_{008}{*}(\text{TeacherMathMajor})_k + \gamma_{009}{*}(\text{StudentBlack})_k + \gamma_{0010}{*}(\text{StudentWhite})_k + u_{00k}$.

$\beta_{01k} = \gamma_{010}$.

$\beta_{10k} = \gamma_{100}$.

$\beta_{20k} = \gamma_{200}$.

Level 3 variables:

$(\text{CMP2})_k$ – intervention indicator that takes a value of 1 for an intervention school and 0 for a control school.

$(\text{PretestSch})_k$ – average pretest score for students in school *k*, grand-mean centered.

$(\text{Juris1})_k$ – jurisdiction indicator variable that takes a value of 1 for jurisdiction 1 and 0 for jurisdictions 2, 3, and 4, grand-mean centered.

$(\text{Juris2})_k$ – jurisdiction indicator variable that takes a value of 1 for jurisdiction 2 and 0 for jurisdictions 1, 3, and 4, grand-mean centered.

$(\text{Juris3})_k$ – jurisdiction indicator variable that takes a value of 1 for jurisdiction 3 and 0 for jurisdictions 1, 2, and 4, grand-mean centered.

$(\text{SchoolUrban})_k$ – school locale variable that takes a value of 1 for urban and 0 for rural, suburban, and small city, grand-mean centered.

$(\text{TeacherWhite})_k$ – percentage of teachers in school *k* that are White, grand-mean centered.

$(\text{TeacherMathMajor})_k$ – percentage of teachers in school *k* that majored in mathematics during college, grand-mean centered.

$(\text{StudentBlack})_k$ – percentage of students in school *k* that are Black, grand-mean centered.

$(\text{StudentWhite})_k$ – percentage of students in school *k* that are White, grand-mean centered.

Level 3 coefficients:

γ_{000} – *adjusted* average student outcome across all control schools, when CMP2 = 0.

γ_{001} – *adjusted* difference between intervention and control schools (the intervention's effect) on student outcome.

γ_{002} – adjusted relationship between school average pretest and school average student outcome.

γ_{003} – adjusted average difference in student outcome between jurisdiction 1 and jurisdiction 4.

γ_{004} – adjusted average difference in student outcome between jurisdiction 2 and jurisdiction 4.

γ_{005} – adjusted average difference in student outcome between jurisdiction 3 and jurisdiction 4.

γ_{006} – adjusted average difference in student outcome between urban and non-urban schools.

γ_{007} – adjusted increase in school average TerraNova score for every unit increase in the percentage of teachers within school who are White.

γ_{008} – adjusted increase in school average TerraNova score for every unit increase in the percentage of teachers within school who majored in mathematics during college.

γ_{009} – adjusted increase in school average TerraNova score for every unit increase in the percentage of students within school who are Black.

γ_{0010} – adjusted increase in school average TerraNova score for every unit increase in the percentage of students within school who are White.

γ_{010} – adjusted average relationship between class average pretest and class average student outcome across all schools.

γ_{100} – adjusted average relationship between student pretest and student outcome across all schools.

γ_{200} – adjusted average difference in outcome between missing and not missing pretest across all schools.

u_{00k} – a random error associated with school *k* on school average student outcome with $u_{00k} \sim N(0, \tau_\beta)$.

Of primary interest among the level-3 coefficients is γ_{001}, which represents CMP2's adjusted main effect on student mathematics achievement as measured by the TerraNova. A statistically significant value of γ_{001} will allow rejection of the null hypothesis that students in intervention schools are not different in the mathematics achievement outcome than their counterparts in control schools. For the sensitivity analysis, this coefficient was compared with that produced by the two-level main model to determine how sensitive the impact estimate generated by the main HLM is to that generated by the three-level model.

Appendix H. Implementation Analysis for Intervention and Control Schools

To measure fidelity of intervention implementation, data were collected through the monthly online survey for intervention teachers using CMP2 and through classroom observations in intervention and control classrooms (table H1).

Table H1. Benchmarks for CMP2 implementation

Data source	Items on the data collection instrument	Aggregate variable	Fidelity benchmarks
1a. Curriculum used by intervention teachers			
Monthly online survey	Summary of responses to question 1: Name of CMP2 units completed over the year	Number of books completed in impact year	Publisher guidelines: 6 of 8 units
	Question 4: The amount of time spent per week on CMP2	Average time spent on mathematics	Developer guidelines: 50+ minutes per day or 250+ minutes per week
	Questions 4 and 5: CMP2 is the primary curriculum being used	Qualitative assessment	CMP2 is the primary curriculum being used with some supplementation
1b. Curriculum used by control teachers			
Classroom observation protocol	Item 20	Textbook used in the control class section	Use of any CMP or CMP2 materials
2. Instructional practices (Key instructional practices according to the CMP2 publisher's guidelines). Data collected for intervention and control teachers and compared.			
Classroom observation protocol	Items 5, 15, 16a, 16d, 16e, 16f, 16g	Teacher factors related to student responsibility for learning and complex thinking	There is no fidelity benchmark for these practices
	Items 10a–10e	Student evidence of responsibility for learning and complex thinking	
	Items 16h, 16i, 16k, 16l, 22	Making connections	
	Items 7c–7e	Percentage of time on CMP2-like activities	
	Items 7b, 7f	Percentage of time on activities less like CMP2	

Source: Authors' analysis of study notes.

The data gathered through classroom observations was coded to create indicators for analyzing implementation to compare instruction in the intervention and control schools (table H2).

Table H2. Indicators from observation protocol used to examine implementation and compare instruction in intervention and control schools

Instruction component and subcomponents	Criterion for scoring using the site-visit protocol	Coding instruction
Making connections		# out of 5 points for each teacher used to compute average teacher points for this area of practice.
16h: Connect concepts taught in class to things students already know	The teacher connects the new concepts to something the students learned previously.	Yes = 1 No = 0
16i: Connect concepts taught in class to the real world	The teacher connects something in the mathematics lesson or investigation to a situation in the real world that the students can relate to.	Yes = 1 No = 0
16k: Use alternative teaching strategies if students fail to understand the lesson	The teacher uses graphic organizers, scaffolding questions, or other methods of working with students who do not understand the lesson.	Yes = 1 No = 0
16l: Assess student's prior knowledge of a concept	The teacher asks questions to assess student's prior knowledge and understanding of the concepts to be covered that class period.	Yes = 1 No = 0
22: Prelaunch reference to real-world connection	The teacher uses a brief description of the real-world application to launch the lesson.	Yes = 1 No = 0
Teacher factors related to student responsibility for learning and complex thinking		# out of 11 for each teacher used to compute average teacher points for this area of practice.
16g: Expect students to engage in complex thinking	The teacher requires that students justify their thinking, explain their reasoning; the teacher does not accept just an answer. Students also may come up with ways to change the problem and explore what might change if that part of the problem is changed.	Yes = 1 No = 0
5: Classroom seating conducive to group work and/or pair work.	Students seated in clusters of 2–5 students and given opportunity to work together.	1 point given for groupwork or pair work seating; 0 for the rest
15: Pedagogy ("Sage on the stage vs. "Guide on the Side")	The Sage on the Stage teacher (1 point) does not always have to be at the board; he or she could be walking around the room, but it would be clear that the teacher is the one with the information and that his or her method (procedure) is the best/most efficient one or the one the students are being asked to use. To be a Guide on the Side teacher (5 points), the teacher elicits the information from the students and allows the students to make the connections.	1–5 points given per teacher for pedagogy.

Instruction component and subcomponents	Criterion for scoring using the site-visit protocol	Coding instruction
16a: Present student learning goals related to the lesson or activity	This should be related to mathematics. It can be informal or formal, where the teacher explains to the students the mathematical goals for the lesson.	Yes = 1 No = 0
16d: Students answering each other's questions	The teacher does not have to formally invite students to answer each other's questions. If students are answering each other's questions, it is part of the classroom environment that the teacher may have set up during another class.	Yes = 1 No = 0
16e: Allow students to work with and help each other	The teacher sets up the class to allow students to work with each other in pairs or groups, or to help each other as needed this item can be checked.	Yes = 1 No = 0
16f: Encourage curiosity and creativity in students	This is often seen when the teacher provides or allows different materials or methods for the students to use to explore the problem. The students could ask about whether a problem could be solved another way.	Yes = 1 No = 0
Student evidence of responsibility for learning and complex thinking: Class discussion		# out of 5 points for each teacher used to compute average teacher points for this area of practice.
10a: Students answer each other's questions.	In a class discussion, a student asks a question, but the teacher allows another student to answer the question.	Yes = 1 No = 0
10b: Students make connections to previous lessons.	The students are making connections. The students would need to connect to something they learned in another lesson.	Yes = 1 No = 0
10c: Students introduce more than one way to approach a problem.	During a discussion, one student would share his or her thinking of a problem, and then another student would share a strategy that is different.	Yes = 1 No = 0
10d: Students take turns answering teacher probes.	In a class discussion, multiple students are responding to teacher's questions.	Yes = 1 No = 0
10e: Students collaborate to solve a problem.	In a class discussion, a problem could be proposed and different students could share their ideas as to how to start solving it, collaborating with the other students to reach a solution.	Yes = 1 No = 0

Instruction component and subcomponents	Criterion for scoring using the site-visit protocol	Coding instruction
Student evidence of responsibility for learning and complex thinking: groupwork and pair work		# out of 5 points for each teacher used to compute average teacher points for this area of practice
10a: Students answer each other's questions.	A student asks a question, but the teacher allows another student to answer the question. This typically occurs naturally in groupwork and pair work, but not always, so observe the groups/pairs to see if this is happening.	If either pair work or groupwork have a Yes, then given a 1. Both No is scored a 0.
10b: Students make connections to previous lessons.	Students would need to connect to something they learned in another lesson.	If either pair- or groupwork have a Yes, then given a 1. Both No is scored a 0.
10c: Students introduce more than one way to approach a problem.	During pair work or groupwork, one student would share his/her thinking of a problem, and then another student would share a strategy that is different.	If pair- or groupwork have a Yes, then given a 1. Both No is scored a 0.
10d: Students take turns answering teacher probes.	In groupwork, the teacher comes to the group and asks a question, and more than just one student responds.	If pair- or groupwork have a Yes, then given a 1. Both No is scored a 0.
10e: Students collaborate to solve a problem.	A problem could be proposed and different students could share their ideas as to how one might start solving it, collaborating with the other students to reach a solution.	If pair- or groupwork have a Yes, then given a 1. Both No is scored a 0.
Percentage of time on activities like CMP2		Average for each teacher of these 3 activities, averaged across teachers
Class discussion	The teacher may do some telling of information by connecting the student's questions and observations to the mathematics goals of the lesson. The students should be involved by asking/answering questions, sharing observations, and the like.	Percentage of total class time
Small groupwork	This can happen in many ways, such as students independently read and solve the first problem for 2 minutes, students turn to a partner and share their answers and justify their thinking for 4 minutes, and partners join 2 other partner pairs to make groups of 6 to solve the main investigation for 15 minutes.	Percentage of total class time
Pair work	Students given opportunity to work with another person and share their ideas.	Percentage of total class time

Instruction component and subcomponents	Criterion for scoring using the site-visit protocol	Coding instruction
Percentage of time on activities that are less like CMP2		Average for each teacher of these 3 activities, then averaged across teachers
Lecture	The teacher is telling students information that they need to learn, similar to a college lecture. The teacher does not have to physically be at the board. He or she can be standing at the back of the room, walking around the room. The key is that the teacher is the one holding the knowledge or information and his/her role is to transmit that information into the student's minds. The student's role is to listen and/or record the information in notes.	Percentage of total class time
Independent work	Students independently work on their assignments.	Percentage of total class time

Source: Study observation protocols.

An HLM was used to correct the standard error of the mean difference for clustering in a comparison of intervention teachers to control teachers on instructional practices, because teachers were nested within schools. This HLM is described below:

$$\text{TeacherOutcome}_{ij} = \beta_{00} + r_{ij}$$

$$\beta_{00} = \gamma_{00} + \gamma_{01}{*}\text{CMP2}_j + u_{0j}$$

CMP2, where 1 = teacher in intervention school and 0 = teacher in control school, uncentered.

The above HLM was used to estimate the mean difference for intervention and control teachers for each instructional practice component (see table F2). The results are based on the impact year observation data for the control and intervention teachers for fall 2009 and spring 2010 (see chapter 3).

Appendix I. Cost of the Curriculum and Professional Development

Cost information for the intervention was obtained from a line-item spreadsheet completed by the PD provider and a cost elements survey. The cost of the intervention involves two days of summer PD to introduce the new curriculum, three days of PD during the school year, and provision of curricula materials.

Table I1 summarizes the costs of these components for the first year of the intervention. The total cost represents what it would cost a district with a similar number of teachers and students to purchase the curriculum and PD from the publisher.

Table I1. Cost of PD and curriculum for the implementation year

PD cost component	Unit	Number of units	Total cost	Percentage of total cost
Grand total			$503,411	100.0
PD trainer fees	Training	29	$40,500	8.0
PD for teachers				
Teacher stipend	Teacher	92	$37,861	7.5
Substitute cost	Day	210	$33,600	6.7
Teacher transportation	Teacher	10	$594	0.1
Materials				
Teacher versions, training materials	Teacher	105	$66,934	13.3
Student materials	Student	4,059	$323,923	64.3

Source: Cost charged by the publisher based on the needs of the sample.

Several components of the PD for the CMP2 curriculum package are shown in the cost breakdown in table I1 for the implementation year. The publisher's PD trainers provided the five days of PD for $40,500 for the implementation year. A stipend was provided to teachers for attending PD outside their regular workday, which totaled $37,861 across the intervention schools for both the intervention year and for new teachers trained during the impact year. Teachers attending PD during a regular workday were provided with a substitute, paid for at an average of $160 per day, resulting in an overall cost of $33,600 for substitutes. To allow for teachers to work in small groups in a collaborative learning environment for PD, small schools with one or two teachers participated in PD with other schools in the study. Transportation costs were paid for any teachers needing to attend such PD outside their school. This resulted in $594 for transportation reimbursements. The cost of the teacher editions and training materials for 75 teachers totaled $66,934. The cost for the student materials for the 3,038 students in the intervention was $323,923.

In preparation for the impact year, schools were provided with additional teacher and student materials per their request. In some cases, an increase in enrollment was the reason for the need for additional materials. In others, some of the student materials had been lost or damaged. The publisher's trainers provided two days of PD to teachers who

were new to teaching CMP2 during the summer prior to the impact year. Table I2 provides the cost breakdown of the PD and the curriculum for the impact year.

Table I2. Cost of PD and curriculum for the impact year

PD cost component	Unit	Number of units	Total cost	Percentage of total cost
Grand total			$30,472	100.0
PD trainer fees	Trainer	0	$0	0
PD for teachers				
Teacher stipend	Teacher	21	$7,973	26.2
Substitute cost	Day	0	$0	0
Teacher transportation	Teacher	14	$1,024	3.4
Materials				
Teacher versions, training materials	Teacher	4	$3,275	10.7
Student materials	Student	1,042	$18,200	59.7

Source: Cost charged by the publisher based on the needs of the sample.

The average cost of the CMP2 intervention is approximately $14,383 per school, $4,794 per teacher, and $124 per student for one year (table I3).

Table I3. Cost per school, teacher, and student associated with CMP2 implementation

Unit	Intervention group *n*	Cost per unit
School	35	$14,383
Teacher	105	$4,794
Student	4,059	$124

Note: Total cost = $503,411.
Source: Cost charged by the publisher based on the needs of the sample.

The cost information in table I3 does not include any costs for administrator time. It is not clear whether administrators had to spend more time than usual to ensure proper implementation of CMP2. Data were not collected on administrative time spent on implementation of CMP2. If administrators spent extra time implementing the curriculum, additional costs might have been involved. There were almost surely some costs that were not measured and reported in tables I1and I2. No cost was reported for facility rentals for PD during the intervention. In other settings, costs might be incurred for facility use.

Appendix J. Results from Hierarchical Linear Models to Estimate the Impact of CMP2

Table J1. Models estimated to quantify the impact of CMP2 on student TerraNova posttest scores

		Main model (*n* = 5,677)			Unadjusted model (*n* = 5,677)			Urban Only model (*n* = 5,677)			Additional covariates model (*n* = 5,677)		
Parameter		**Estimate**	**SE**	***p***	**Estimate**	**SE**	***p***	**Estimate**	**SE**	***p***	**Estimate**	**SE**	***p***
Fixed effects													
Intercept	γ_{00}	682.16	1.49	.000	690.17	3.71	.000	686.51	3.11	.000	682.57	1.52	.000
CMP2	γ_{01}	0.60	2.12	.777	–13.51	5.08	.010	–6.87	4.32	.117	–0.17	2.19	.938
School mean pretest	γ_{02}	0.94	0.10	.000	—	—	—	—	—	—	0.96	0.10	.000
School Juris1	γ_{03}	15.17	11.67	.199	—	—	—	—	—	—	12.23	12.06	.316
School Juris2	γ_{04}	–2.72	7.74	.727	—	—	—	—	—	—	–3.63	7.81	.644
School Juris3	γ_{05}	-0.31	2.80	.914	—	—	—	—	—	—	–0.56	2.87	.845
School: Urban	γ_{06}	4.54	3.84	.243	—	—	—	–27.68	4.97	.000	5.32	3.88	.176
Teacher: White	γ_{07}	0.28	4.60	.952	—	—	—	—	—	—	–11.53	14.65	.435
Teacher: Math major	γ_{08}	3.80	4.72	.425	—	—	—	—	—	—	3.35	4.74	.483
Student: Black	γ_{09}	–18.01	9.24	.056				—	—	—	–12.57	9.96	.213
Student: White	γ_{010}	3.13	7.27	.668				—	—	—	0.04	8.34	.996
Teacher: Black	γ_{011}	—	—	—	—	—	—	—	—	—	–13.03	14.82	.384
Student: Female	γ_{012}	—	—	—	—	—	—	—	—	—	-0.39	0.36	.277
Student: TerraNova pretest	γ_{10}	0.83	0.01	.000	—	—	—	—	—	—	0.83	0.01	.000
Student: Pretest missing	γ_{20}	14.53	1.67	.000	—	—	—	—	—	—	–14.51	1.67	.000
Variance estimates													
Students (level 1)	r_{ij}	523.81	—	—	1,249.85	—	—	1, 249.98	—	—	523.78	—	—
Schools (level 2)	u_{0j}	47.00	—	.000	390.48	—	.000	253.66	—	.000	47.06	—	.000

Note. The variables in the respective models were centered as described in appendix G.
Source: Analysis of student TerraNova posttest scores.

Table J2. Models estimated to quantify the impact of CMP2 on student TerraNova posttest scores

Parameter label	Main model (*n* = 5,677)			Case deletion model (*n* = 5,475)			Three-level model (*n* = 5,677)[a]		
	Parameter estimate	SE	*p*	Parameter estimate	SE	*p*	Parameter estimate	SE	*p*
Fixed effects									
Intercept	682.16	1.49	.000	682.51	1.40	.000	680.62	1.43	.000
CMP2	0.60	2.12	.777	0.88	2.00	.662	1.05	2.03	.609
School mean pretest	0.94	0.10	.000	0.93	0.09	.000	0.94	0.09	.000
School Juris1	15.17	11.67	.199	14.19	11.07	.206	–15.70	13.26	.242
School Juris2	–2.72	7.74	.727	–3.98	7.29	.587	–14.34	11.98	.237
School Juris3	–0.31	2.80	.914	–0.01	2.65	.998	–13.63	11.99	.261
School: Urban	4.54	3.84	.243	3.26	3.63	.373	5.74	3.69	.126
Teacher: White	0.28	4.60	.952	0.21	4.37	.962	0.09	4.52	.983
Teacher: Math major	3.80	4.72	.425	4.18	4.45	.352	3.30	4.59	.475
Student: Black	–18.01	9.24	.056	–18.02	8.69	.043	–12.26	8.84	.171
Student: White	–3.13	7.27	.668	–5.77	6.86	.405	1.85	6.91	.790
Class section: TerraNova Pretest[b]	—	—	—	—	—	—	1.01	0.03	.000
Student: TerraNova pretest	0.83	0.01	.000	0.84	0.01	.000	0.78	0.01	.000
Student: Pretest missing	–14.53	1.67	.000	—	—	—	–13.52	1.66	.000
Variance estimates									
Students (level 1)	523.81	—	—	478.31	—	—	493.92	—	—
Classes (level 2)	47.00	—	.000	41.27	—	.000	23.62	—	.000
Schools (level 3)	—	—	—	—	—	—	36.04	—	.000

a. The level 2 (classroom) sample size was *n* = 310.
b. This parameter is not specified for the two-level models but is specified in the three-level model.
Note: The variables in the respective models were centered as described in appendix G.
Source: Analysis of student TerraNova posttest scores.

Table J3. Models estimated to quantify the impact of CMP2 on student PTV posttest scores

Parameters		**Main model** (***n***=5,584)			**Unadjusted model** (***n*** =5,584)			**Case deletion model** (***n*** = 5,043)			**Additional covariates model** (***n*** = 5,584)		
		Parameter estimate	**SE**	***P***	**Parameter estimate**	**SE**	***p***	**Parameter estimate**	**SE**	***p***	**Parameter estimate**	**SE**	***p***
Fixed effects													
Intercept	γ_{00}	36.67	0.29	.000	36.71	0.26	.000	36.66	0.28	.000	36.74	0.30	.000
CMP2	γ_{01}	0.65	0.40	.109	0.59	0.35	.102	0.61	0.40	.133	0.53	0.42	.221
School mean pretest	γ_{02}	–0.11	0.15	.448	—	—	—	–0.15	0.14	.270	–0.10	0.15	.508
School Juris1	γ_{03}	–1.29	3.00	.668	—	—	—	–1.12	2.97	.708	–1.92	3.07	.533
School Juris2	γ_{04}	–2.34	1.46	.114	—	—	—	–2.16	1.46	.144	–2.58	1.50	.090
School Juris3	γ_{05}	0.33	0.54	.542	—	—	—	0.40	0.53	.462	0.30	0.56	.589
School: Urban	γ_{06}	–0.08	0.76	.916	—	—	—	–0.25	0.76	.742	–0.01	0.78	.991
Teacher: White	γ_{07}	–0.64	0.94	.497	—	—	—	–0.61	0.94	.521	0.13	2.85	.965
Teacher: Math major	γ_{08}	0.28	0.94	.763	—	—	—	0.36	0.93	.701	0.21	0.96	.830
Student: Black	γ_{09}	–0.79	1.75	.655	—	—	—	–0.80	1.74	.648	–0.56	1.93	.773
Student: White	γ_{010}	–0.99	1.57	.529	—	—	—	–1.37	1.54	.380	–1.03	1.75	.558
Teacher: Black	γ_{011}	—	—	—	—	—	—	—	—	—	0.69	2.84	.808
Student: Female	γ_{012}	—	—	—	—	—	—	—	—	—	–0.08	0.07	.258
Student: TerraNova pretest	γ_{10}	0.02	0.02	.291	—	—	—	0.02	0.02	.270	0.02	0.02	.291
Student: Pretest missing	γ_{20}	0.35	0.33	.296	—	—	—	—	—	—	0.36	0.33	.285
Variance estimates													
Students (level 1)	r_{ij}	52.54	—	—	52.54	—	—	52.65	—	—	52.53	—	—
Schools (level 2)	u_{0j}	1.25	—	.000	1.14	—	.000	1.15	—	.000	1.32	—	.000

Note: The variables in the respective models were centered as described in appendix G
Source: Analysis of student PTV posttest scores.

References

Adelman, C. (2006). *The toolbox revisited: Paths to degree completion from high school through college*. Washington, DC: Office of Vocational and Adult Education.

Augustine, N. R., Vagelos, P. R., & Wulf, W. A. (2005, October). *Rising above the gathering storm: Energizing and employing America for a brighter economic future.* National Academies Statements on Science, Technology, and Global Economic Competitiveness to the 109th Congress (First Session), Washington, DC.

Bailey, T. (2009). Challenge and opportunity: Rethinking the role and function of developmental education in community college. *New Directions for Community Colleges, 2009*(145), 11–30.

Battista, M. T. (1999). The mathematical miseducation of America's Youth: Ignoring research and scientific study in education. *Phi Delta Kappan, 80,* 424–433.

Betts, J. R., & Rose, H. (2001). *Math matters: The links between high school curriculum, college graduation and earnings*. San Francisco: Public Policy Institute of California.

Bishop, A. J. (1988). Mathematics education in its cultural context. *Educational Studies in Mathematics, 19*(2), 179–191.

Bloom, H. S. (Ed.) (2005). *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage.

Bloom, H. S., Richburg-Hayes, L., and Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Boaler, J. (2002). Learning from teaching: Exploring the relationship between reform curriculum and equity. *Journal for Research in Mathematics Education, 33,* 239–258.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 445–455). Newbury Park, CA: Sage.

Burghardt, J., Deke, J., Kisker, E., Puma, M., & Schochet, P. (2009). *Regional Educational Laboratory rigorous applied research studies: Frequently asked questions* (MPR Reference No.: 6300-040). Washington, DC: Mathematica Policy Research.

Business-Higher Education Forum. (2007). *An American imperative: Transforming the recruitment, retention, and renewal of our nation's mathematics and science teaching workforce*. (ERIC Document Reproduction Service No. ED503709)

Chung, I. (2004). A comparative assessment of constructivist & traditionalist approaches to establishing mathematical connections in learning multiplication. *Education, 125,* 271–278.

Cobb, P. (1989). Experiential, cognitive, and anthropological perspectives in mathematics education. *For the Learning of Mathematics, 9*(2), 32–42.

Cobb, P., Yackel, E., & Wood, T. (1992). A constructivist alternative to the representational view of mind in mathematics education. *Journal for Research in Mathematics Education, 23*(1), 2–33.

CTB/McGraw-Hill. (2003). *CAT^{TM} Technical Report: TerraNova* (2nd ed.). Monterey, CA.

Eccles, J. S., & Jacobs, J. E. (1986). Social forces shape math attitudes and performance. *Signs: Journal of Women in Culture and Society, 11,* 367–380.

Eccles, J., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents. achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215–225.

Eddy, R. M., Berry, T., Aguirre, N., Wahlstrand, G., Ruitman, H., & Mahajan, N. (2008). *The effects of Connected Mathematics Project 2 on student performance: Year one final report.* Claremont, CA: Claremont Graduate University, Institute of Organizational and Program Evaluation Research.

Elbers, E. (2003). Classroom interaction as reflection: Learning and teaching mathematics in a community of inquiry (Abstract). *Educational Studies in Mathematics, 54*(1), 77–100.

Elledge, A., Le Floch, K.C., Taylor, J., and Anderson, L. (2009). *State and Local Implementation of the No Child Left Behind Act. Volume V*: *Implementation of the 1 Percent Rule and 2 Percent Interim Policy Options.* Washington, DC: U.S. Department of Education.

Erlwanger, S. (1973). Benny's conception of rules and answers in IPI mathematics. *Journal of Children's Mathematical Behavior, 1*(3), 157–283.

Flay, B. R. (1986). Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Preventive Medicine, 15,* 451–474.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., Koscicki, E. K., Schinke, S., Valentine, J. C., & Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science, 6*(3), 151–175.

Friedman, L. M., Furberg, C. D., and DeMets, D. L. (1997). *Fundamentals of clinical trials* (3rd ed.). New York: Springer.

Goals 2000: Educate America Act of 1994, P. L. 103–227, U. S. C. 5801 *et seq*.

Greene, J. P., & Forster, G. (2003). *Public high school graduation and college readiness rates in the United States.* New York: Center for Civic Innovation at the Manhattan Institute.

Grubb, N. W., and Cox, R. D. (2005). Pedagogical alignment and curricular consistency: The challenges for developmental education. *New Directions for Community Colleges, 2005*(129), 93–103.

Helmke, A. (1989). Incentive value of success and failure in school. Developmental trends and impact on academic achievement. In F. Halisch & J. J. L. van den Berchen (Eds.), *International perspectives on achievement and task motivation* (pp. 225–237). Lisse, The Netherlands: Swets & Zeitlinger.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177–189.

Hodge, L. L., Zhao, Q., Visnovska, J., & Cobb, P. (2007). Revisiting discourse as an instructional resource: Practices that create spaces for learning and student contributions. In J. H. Woo, H. C. Lew, K. S. Park, & D. Y. Seo (Eds.), *Proceedings of the 31st International Conference for the Psychology of Mathematics Education* (Vol. 3, pp. 41–48). Seoul: PME.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Joyce, B., & Showers, B. (1995). *Student achievement through staff development* (2nd Ed.). New York: Longman.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S., & Phillips, E. D. (2006a). *Connected Mathematics 2 implementing and teaching guide.* Boston: Pearson Prentice Hall.

Lappan, G., Fey, J. T., Fitzgerald, W. M., Friel, S., & Phillips, E. D. (2006b). *Teacher's guide for the Connected Mathematics 2 prime time unit*. Boston: Pearson Prentice Hall.

Lee, J., Grigg, W. S., & Dion, G. S. (2007). *The nation's report card: Mathematics 2007* (NCES 2007–494). Washington, DC: National Center for Education Statistics.

Lucas, C. A. (2006). Is subject matter knowledge affected by experience? The case of composition of functions. In J. Novotná, H. Moraová, M. Krátká, & N. Stehliková (Eds.), *Proceedings of the 30th International Conference for the Psychology of Mathematics Education* (Vol. 4, pp. 97–104). Prague: PME.

Martin, D. B. (2007). Beyond missionaries or cannibals: Who Should Teach mathematics to African-American children? *The High School Journal, 91*(1), 6–28.

McGraw-Hill. (2002). *TerraNova, 2nd edition: Technical bulletin 1*. Monterey, CA.

Meaney, T. (2007). Weighing up the influence of context on judgments of mathematical literacy. *International Journal of Science and Mathematics Education,* 5(4), 681–704.

Moss, J., & Beatty, R. (2006). Knowledge building and knowledge forum: Grade 4 students collaborate to solve linear generalizing problems. In J. Novotná, H. Moraová, M. Krátká, & N. Stehliková (Eds.), *Proceedings of the 30th International Conference for the Psychology of Mathematics Education* (Vol. 4, pp. 193–199). Prague: PME.

Mullis, I. V. S., Martin, M. O., & Foy, P. (2008). *TIMSS 2007 international mathematics report*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center.

National Center for Education Evaluation and Regional Assistance. (2011). *Assessment of regional educational needs: Reports from ten regional educational laboratories and statistical data*. Retrieved July 28, 2011, from https://www.fbo.gov/index?s=opportunity&mode=form&id=79d5d6b34726cf9163fbcba3ae468dd8&tab=documents&tabmode=list

National Center for Education Statistics. (n.d.a). *Common Core of Data, 2006–2007*. Public Elementary/Secondary School Universe Survey Data. Washington, DC: U.S. Department of Education, Institute for Education Sciences. Retrieved January 15, 2010, from http://nces.ed.gov/ccd/

National Center for Education Statistics. (n.d.b). *Common Core of Data, 2008–2009*. Public Elementary/Secondary School Universe Survey Data. Washington, DC: U.S. Department of Education, Institute for Education Sciences. Retrieved June 10, 2010, from http://nces.ed.gov/ccd/

National Center for Education Statistics. (2009). *The nation's report card: Mathematics 2009, National Assessment of Educational Progress at grades 4 and 8* (NCES 2010–451). Washington, DC: U.S. Department of Education. Retrieved May 15, 2010, from http://nces.ed.gov/nationsreportcard/pdf/main2009/2010451.pdf

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform: An open letter to the American people.* Washington, DC.

National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA.

No Child Left Behind Act of 2001. (2002). Pub. L. No. 107–110.

Ozmantar, M. F., and Monaghan, J. (2006). Abstraction, scaffolding, emergent goals. In J. Novotná, H. Moraová, M. Krátká, & N. Stehliková (Eds.), *Proceedings of the 30th International Conference for the Psychology of Mathematics Education,* (Vol. 4, pp. 305–312). Prague: PME.

Piantadosi, S. (2005). *Clinical trials: A methodologic perspective.* Hoboken, NJ: Wiley & Sons.

Popkewitz, T. (2004). The alchemy of the mathematics curriculum: Inscriptions and the fabrication of the child. *American Educational Research Journal, 41*(1), 3–34.

Powell, A. B. (2006). Socially emergent cognition: Particular outcome of student-to-student discursive interaction during mathematical problem solving. *Horizontes, 24*(1), 33–42.

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009–0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved August 15, 2010, from http://ies.ed.gov/ncee/pdf/20090049.pdf

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2005). *Strategies for improving precision in group-randomized experiments.* New York: William T. Grant Foundation.

Remillard, J. T., Herbel-Eisenmann, B. A., & Lloyd, G. M. (Eds.). (2009). *Mathematics teachers at work: Connecting curriculum materials and classroom instruction*. New York: Routledge.

Reynolds, A. J., & Walberg, H. J. (1991). A structural model of science achievement. *Journal of Education Psychology, 83*(2), 97–107.

Reynolds, A. J., & Walberg, H. J. (1992). A structural model of science achievement and attitude: An extension to high school. *Journal of Educational Psychology, 84*(3), 371–382.

Ridgway, J., Zawojewski, S., Hoover, M. V., & Lambdin, D. V. (2003). Student attainment in the Connected Mathematics curriculum. In S. Senk & D. R. Thompson (Eds.), *Standards based school mathematics curricula: What are they? What do students learn?* (pp. 193–224). Hillsdale, NJ: Erlbaum.

Ridlon, C. (2004). The effect of a problem centered approach to mathematics on low-achieving 6th-graders. *Focus on Learning Problems in Mathematics, 26*(2), 18–48.

Riordan, J., & Noyce, P. (2001). The impact of two standards-based mathematics curricula on student achievement in Massachusetts. *Journal for Research in Mathematics Education, 32*, 368–398.

Schneider, C. (2000). Connected Mathematics and the Texas Assessment of Academic Skills. (Doctoral dissertation, University of Texas at Austin). *Dissertation Abstracts International, 62,* 2.

Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs.* Princeton, NJ: Mathematica Policy Research.

Schoen, H. L., and Hirsch, C. R. (2003). Responding to calls for change in high school mathematics: Implications for collegiate mathematics. *The American Mathematical Monthly, 110*(2), 109–123.

Schwab, J. (1975). Learning community. *Center Magazine, 8*(3), 30–44.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi experimental designs for generalized causal inference.* Boston: Houghton Mifflin.

Steen, L. A. (2007). How mathematics counts. *Educational Leadership, 65*(3), 8–14.

Stigler, J. W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS videotape classroom study: Methods and findings from an exploratory research project on eighth-grade mathematics instruction in Germany, Japan, and the United States* (NCES 1999–074). Washington, DC: U.S. Government Printing Office.

Suarez, T. M., Torlone, T. J., McGrath, S. T., & Clark, D. L. (1991). *Enhancing effective instructional time: A review of research.* (Policy Brief Vol. 1, No. 2.) Chapel Hill, NC, North Carolina Educational Policy Research Center.

Tsao, Y. (2004). A comparison of American and Taiwanese students: Their math perception. *Journal for Instructional Psychology, 31*(3), 206–213.

Useem, E. L. (1992). Getting on the Fast Track in Mathematics: School Organizational Influences on Math Track Assignment. *American Journal of Education, 100*(3), 325–353.

What Works Clearinghouse (2010). Connected Mathematics Project: WWC intervention report. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved May 1, 2007, from http://ies.ed.gov/ncee/wwc/pdf/wwc_cmp_012610.pdf

Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation, University of California, Los Angeles.